Intelligent Systems, Control and Automation:
Science and Engineering

Maria Isabel Aldinhas Ferreira
Joao Silva Sequeira
Mohammad Osman Tokhi · Endre E. Kadar
Gurvinder Singh Virk *Editors*

A World with Robots

International Conference on Robot Ethics: ICRE 2015

Springer

Intelligent Systems, Control and Automation: Science and Engineering

Volume 84

More information about this series at http://www.springer.com/series/6259

Maria Isabel Aldinhas Ferreira
Joao Silva Sequeira · Mohammad Osman Tokhi
Endre E. Kadar · Gurvinder Singh Virk
Editors

A World with Robots

International Conference on Robot Ethics: ICRE 2015

VNIVERSITAS OLISIPONENSIS
U LISBOA | UNIVERSIDADE DE LISBOA
FLUL LETRAS LISBOA
Springer

Editors
Maria Isabel Aldinhas Ferreira
University of Lisbon Center of Philosophy
Lisbon
Portugal

Joao Silva Sequeira
Instituto Superior Técnico
Technical University of Lisbon
Lisbon
Portugal

Mohammad Osman Tokhi
London South Bank University
London
UK

Endre E. Kadar
Department of Psychology
University of Portsmouth
Portsmouth
UK

Gurvinder Singh Virk
InnotecUK Ltd
Cambridge
UK

ISSN 2213-8986 ISSN 2213-8994 (electronic)
Intelligent Systems, Control and Automation: Science and Engineering
ISBN 978-3-319-46665-1 ISBN 978-3-319-46667-5 (eBook)
DOI 10.1007/978-3-319-46667-5

Library of Congress Control Number: 2016953858

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The increasing deployment of robotic technology in many domains of human life will have a substantial impact on the economic, social and cultural tissues of our societies. Though one can already anticipate some of its huge benefits, it also urges us to try to reflect on its impact on fundamental instances of everyday life and also envisage to what extent essential societal values on which we have based our cultures and legal systems may be eventually affected.

On the verge of this technological revolution, experts from academia, industry, military and civilian sectors gathered in the International Conference on Robot Ethics (ICRE 2015)[1] in order to reflect and discuss the main ethical problems resulting from the widespread adoption of robotics. The present book comprehends not only the shared doubts and concerns but above all the common effort to open up new pathways to a future with robots that contributes to a better world.

The book is divided into two parts: Part I presents selected contributions of the main speakers and also those of invited guests. These are organized according to relevant domains they are addressing comprising general normative and ethical issues, the ethics of some robotic applications, social and service robotics, robots in defence and war scenario and, finally, legal issues.

Part II contains the reflections and accounts of the two other events organized during ICRE 2015: a cinema cycle—The Robot Steps in; and the exhibition Nós e os Robots/Os Robots e Nós[2]

Part I starts with Chap. 1. Here, Malle, Scheutz and Austerweil point out that the most ethically challenging role to be played by robots is that of collaborator and social partner. Proposing that such robots must have the capacity to learn, represent, activate, and apply social and moral norms, the authors offer a theoretical analysis of two parallel questions that are: (i) What constitutes this capacity for norms in humans? (ii) How might we implement them in robots?

[1]Lisbon, 23 and 24 October 2015.

[2]We and the Robots/The Robots and Us, in the English translation.

In Chap. 2, Maaike Harbers, Marieke Peeters, and Mark Neerincx analyse how a robot system's characteristics affect people's perception of its autonomy. Based on a survey aimed at identifying the rate of autonomy assigned by firefighters to a number of search and rescue robots with different shapes and in distinct situations, the authors were able to identify seven distinct aspects of perceived autonomy.

In Chap. 3, Sean Welsh argues that the critical work in deontic reasoning is better done in the knowledge representation rather than reasoning of a normative system. In this chapter, the author describes a way to formalize complex normative decisions using predicate logic and graph databases.

In Chap. 4 by Selmer Bringsjord proposes an ethical hierarchy (EH) that can be applied to both robots and humans. This hierarchy is catalysed by the question: Can robots be more moral than humans? According to Bringsjord, the light shed by EH reveals why an emphasis on legal obligation for robots is inadequate, and why atleast the vast majority of today's state-of-the-art deontic logics are morally inexpressive, whether they are intended to formalize the ethical behaviour of robots or humans.

Chapter 5 by Wilhelm E.J. Klein on Robots and Free Software examines whether the arguments put forward by free software advocated in the context of computers also apply for robots. Summarizing their key arguments Klein explores whether or not they appear transferable to robot use case scenarios. Issues related to robot ethics for children–robot studies reported in contemporary peer-reviewed papers are also presented.

In Chap. 6, Jaeeun Shim and Ronald C. Arkin address the particular benefits brought by robotic technology to the domain of healthcare, namely to patients with Parkinson's disease. The authors point out that since these patients cannot readily communicate their internal and external states due to their limited motor control abilities, they may experience the loss of dignity during therapy with their caregivers. Shim and Arkin postulate that a companion robot can remedy this challenge and reduce the communication gap between the patient and the caregiver smoothing and increasing the effectiveness of the interactions. To achieve this goal, they have developed a robot architecture that can help prevent the loss of dignity in patient–caregiver relationships. The primary goal of this robot mediator is to ensure patients' and caregivers' dignity during their interactions.

Chapter 7 by Harbers, de Greeff, Kruijff-Korbayova, Neerincx, and Hindriks addresses a particular field that, according to the authors, is under-examined when compared to other robotic application areas. The chapter describes the outcomes of several value assessment workshops that were conducted with rescue workers, in the context of a European research project on robot-assisted search and rescue (SAR). These outcomes are analysed, key ethical concerns and dilemmas are identified and recommendations for future ethical-related research was identified leading to responsible development and deployment of SAR robots.

M. Kyriakidou, K. Padda and L. Parry's study on Chap. 8 explore how robot ethics in children–robot interaction studies are described in contemporary peer-reviewed papers. The outcomes of a survey conducted on 27 articles indicate problematic applications of reporting robot ethics in peer-reviewed journals and

highlight the necessity for journals to consider stricter action on this aspect of publication.

In Chap. 9, Sjur Dyrkolbotn considers non-contractual liability for harm caused by artificially intelligent systems and provides a typology of different possible ways to approach the liability issue. The paper argues that the traditional robot-as-tool perspective should be maintained but warns that new techniques need to be developed, at the intersection between computer science and law, to support reasoning about the liability implications when autonomous technologies interact with their environment and cause harm.

Chapter 10 provides fundamental insights into the difficulties of autonomous and mixed vehicle control Endre E. Kadar, Anna Köszeghy and Gurvinder Singh Virk address this problem based on the evidence provided by three case studies.

Chapter 11 by L. Beton, P. Hughes, S. Barker, M. Pilling, L. Fuente and N.T. Crook refers that due largely to the introduction of new technologies such as force sensing, it is now possible to have humans present within the workspace of a robot in an industrial setting. However, the authors emphasize that physical safety is not the only consideration when attempting to develop robots that are truly able to collaborate with humans. The establishment of trust lies at the heart of any such collaboration. The authors argue that trust in a robot depends, at least in part, on perceived safety and perceived intelligence, and that these, in turn, depend on the collaborative strategies that the robot adopts. A significant number of studies have been performed on human–robot collaboration strategies. One of the key areas of interest is in the adoption of leader/follower roles in the collaboration.

Also addressing industrial robotics, S.R. Fletcher and P. Webb, in Chap. 12, claim that technological advances will cause a change in the way industrial robots are viewed and traditionally operated. This means they will leave their usually highly secluded environments being deployed to work more closely and collaboratively with people in monitored manufacturing systems with the widespread introduction of small-scale robots and assistive robotic devices. According to the authors, this will not only transform the way people are expected to work and interact with automation, but will also involve much more data provision and capture for performance monitoring. The chapter discusses the background of these developments and the anticipated ethical issues that are likely to be faced.

In Chap. 13, Sean Welsh claims that lethal decision-making is complex and requires detailed analysis to define what is to be banned or regulated. The chapter proposes an extension of the current “single-loop” analysis to two loops: a policy loop and a firing loop. The aim is to clarify what exactly is meant by meaningful human control of a law and to facilitate wording such as might occur in a Protocol VI to be added to the Convention on Certain Conventional Weapons (CCW).

Chapter 14 by Dores Delfim, Ana Baltazar, Teresa Cabral, Isabel Machado and Paula Gonçalves provides an overview of the safety issues of the Portuguese Military Remotely Piloted Aircraft Systems (RPAS), namely the human error,

integration into regulated common national airspace (considering the rules of air) and the airworthiness certification aspects. The chapter also brings out the safety assessment methodology by addressing its application to Antex-X02 RPAS, a platform under development by the Portuguese Air Force Academy.

Chapter 15, by invited guest, Major-General João Vieira Borges regards the theme of robotics in the military domain from a strategic perspective. Considering the trilogy that strategy comprehends—goals, means and threats—three fundamental topics are approached: (i) the need to work at political, strategical, operational and tactical levels (ii) the role of robots in the new security and defence environment and (iii) the importance of incorporating robots in military education.

Chapter 16, a keynote by R. Gélin, highlights the importance for the social/service robotics designer of being aware of the potential ethical and safety issues that may arise from the development of humanoid robots functioning as companions. After a short description of a possible use case, dedicated to the assistance of an elderly person, the author identifies the main concerns from safety and ethical points of view and proposes ways on how to prevent risks.

In Chap. 17, Isabel Ferreira and João Sequeira highlight that demographic trends reveal a significant world-changing age distribution resulting from increased average longevity and the deep decline in fertility rates. In this framework, the use of robotic technology to guarantee prolonged autonomy of senior citizens and their active ageing is an imperative. The authors point out, however, that the use of robotic technology can never replace fundamental bonds, as those that link parents to their children.

Two papers constitute the second part of this book: Chap. 18, and Chap. 19.

In Chap. 18, Rodrigo Ventura and Isabel Ferreira report their effort to bring robotic technology closer to the lay persons in the general public—an educational effort that, in their opinion, should precede the massive deployment of all information and communication technologies and that becomes particularly needed at the verge of a widespread use of robotic technology. This chapter gives a brief account of the content and organization of the exhibition and of how the public reacted to it.

In his paper Chap. 19, José Manuel Martins addresses the role of fiction, namely the role of the cinema in the construction of prototypical mental representations and changing mentalities.

Lisbon
July 2016

Maria Isabel Aldinhas Ferreira
Joao Silva Sequeira
Mohammad Osman Tokhi
Endre E. Kadar
Gurvinder Singh Virk

Acknowledgments

The Editors would like to thank to the following people and institutions. Without them this project would not have been possible.

- Lisbon University, Portugal, namely the Center of Philosophy and Instituto Superior Técnico, was the academic sponsor institution.
- The CLAWAR Association, UK, encouraged the realization of the event and provided support at multiple levels.
- Ciência Viva, Portugal, at Pavilhão do Conhecimento in Lisbon, provided the fantastic venue and gave full logistics support.
- The industry partners, Aldebaran, France, namely through Rodophe Gelin and Petra Koudelkova-Delimoges, and hiBot Robotics, Japan, through Naho Kitano. Both complemented nicely the academic viewpoints discussed at the event.
- The IADE Creative University, Portugal, through Bruno Nobre and Emília Duarte, handled the event image communication.
- Lisbon City Hall, and Lisbon Tourism, Portugal, sponsored the social programme of the event.

Contents

Part I Selected Contributions

1 Networks of Social and Moral Norms in Human and Robot Agents 3
B.F. Malle, M. Scheutz and J.L. Austerweil

2 Perceived Autonomy of Robots: Effects of Appearance and Context 19
Maaike Harbers, Marieke M.M. Peeters and Mark A. Neerincx

3 Formalizing Complex Normative Decisions with Predicate Logic and Graph Databases 35
Sean Welsh

4 A 21st-Century Ethical Hierarchy for Robots and Persons: $\mathscr{EH}$ 47
Selmer Bringsjord

5 Robots and Free Software 63
Wilhelm E.J. Klein

6 An Intervening Ethical Governor for a Robot Mediator in Patient-Caregiver Relationships 77
Jaeeun Shim and Ronald C. Arkin

7 Exploring the Ethical Landscape of Robot-Assisted Search and Rescue 93
Maaike Harbers, Joachim de Greeff, Ivana Kruijff-Korbayová, Mark A. Neerincx and Koen V. Hindriks

8 Reporting Robot Ethics for Children-Robot Studies in Contemporary Peer Reviewed Papers 109
M. Kyriakidou, K. Padda and L. Parry

9 A Typology of Liability Rules for Robot Harms 119
Sjur Dyrkolbotn

10 Safety and Ethical Concerns in Mixed Human-Robot Control of Vehicles 135
Endre E. Kadar, Anna Köszeghy and Gurvinder Singh Virk

11 Leader-Follower Strategies for Robot-Human Collaboration 145
L. Beton, P. Hughes, S. Barker, M. Pilling, L. Fuente and N.T. Crook

12 Industrial Robot Ethics: The Challenges of Closer Human Collaboration in Future Manufacturing Systems 159
S.R. Fletcher and P. Webb

13 Clarifying the Language of Lethal Autonomy in Military Robots 171
Sean Welsh

14 Safety Issues of the Portuguese Military Remotely Piloted Aircraft Systems 185
Delfim Dores, Ana Baltazar, Teresa Cabral, Isabel Machado and Paula Gonçalves

15 Robots and the Military: A Strategic View 199
João Vieira Borges

16 The Domestic Robot: Ethical and Technical Concerns 207
Rodolphe Gelin

17 Robots in Ageing Societies 217
Maria Isabel Aldinhas Ferreira and João Silva Sequeira

Part II Associated Events

18 Nós e Os Robots/Os Robots e Nós: Insights from an Exhibition 227
Rodrigo Ventura and Maria Isabel Aldinhas Ferreira

19 The Robot Steps In: From Normative to Prospective Ethics 233
José Manuel Martins

Index 239

Contributors

Ronald C. Arkin School of Interactive Computing, Georgia Institute of Technology, Atlanta, Georgia

J.L. Austerweil Department of Psychology, University of Wisconsin, Madison, WI, USA

Ana Baltazar Instituto Universitário Militar, CISDI Researcher, Lisbon, Portugal

S. Barker Oxford Brookes University, Oxford, UK

L. Beton Oxford Brookes University, Oxford, UK

João Vieira Borges Portuguese Military Academy, Lisboa, Portugal

Selmer Bringsjord Rensselaer AI and Reasoning (RAIR) Lab, Department of Computer Science, Department of Cognitive Science, Rensselaer Polytechnic Institute (RPI), Troy, NY, USA

Teresa Cabral Academia da Força Aérea Portuguesa, CIAFA Researcher, Lisbon, Portugal

N.T. Crook Oxford Brookes University, Oxford, UK

Delfim Dores Instituto Universitário Militar, CISDI Researcher, Lisbon, Portugal

Sjur Dyrkolbotn Department of Philosophy and Religious Studies, Utrecht University, Utrecht, The Netherlands

Maria Isabel Aldinhas Ferreira Centre of Philosophy of the University of Lisbon, Faculdade de Letras, University of Lisbon, Lisbon, Portugal

S.R. Fletcher Centre for Structures, Assembly and Intelligent Automation, Cranfield University, Cranfield, UK

L. Fuente Oxford Brookes University, Oxford, UK

Rodolphe Gelin SoftBank Robotics, Paris, France

Joachim de Greeff Delft University of Technology, Delft, The Netherlands

Paula Gonçalves Instituto Universitário Militar, CISDI Researcher, Lisbon, Portugal

Maaike Harbers Delft University of Technology, Delft, The Netherlands

Koen V. Hindriks Delft University of Technology, Delft, The Netherlands

P. Hughes Oxford Brookes University, Oxford, UK

Endre E. Kadar Department of Psychology, Keimyung University, Daegu, Korea; Department of Psychology, University of Portsmouth, Portsmouth, UK

Wilhelm E.J. Klein School of Creative Media, City University of Hong Kong, Hong Kong, China

Anna Köszeghy Department of Psychology, University of Portsmouth, Portsmouth, UK

Ivana Kruijff-Korbayová Language Technology Lab, DFKI, Saarbruecken, Germany

M. Kyriakidou Department of Psychology and Behavioural Science, Coventry University, Coventry, UK

Isabel Machado Instituto Universitário Militar, CISDI Researcher, Lisbon, Portugal

B.F. Malle Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence, USA

José Manuel Martins Department of Philosophy, University of Évora Center of Philosophy of the University of Lisbon, Lisbon, Portugal

Mark A. Neerincx TNO Human Factors, Delft University of Technology, Delft, The Netherlands

K. Padda Department of Psychology and Behavioural Science, Coventry University, Coventry, UK

L. Parry Department of Psychology and Behavioural Science, Coventry University, Coventry, UK

Marieke M.M. Peeters Delft University of Technology, Delft, The Netherlands

M. Pilling Oxford Brookes University, Oxford, UK

M. Scheutz Department of Computer Science, Tufts University, Medford, USA

João Silva Sequeira Instituto Superior Técnico/Institute for Systems and Robotics, Universidade de Lisboa, Lisbon, Portugal

Jaeeun Shim School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia

Rodrigo Ventura Institute for Systems and Robotics, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal

Gurvinder Singh Virk InnotecUK Ltd, Cambridge, UK

P. Webb Centre for Structures, Assembly and Intelligent Automation, Cranfield University, Cranfield, UK

Sean Welsh Department of Philosophy, University of Canterbury, Christchurch, New Zealand

Part I
Selected Contributions

Right now the major challenge for even thinking about how robots might be able to understand moral norms is that we don't understand on the human side how humans represent and reason if possible with moral norms.

Mathias Scheutz, "How to Build a Moral Robot", Spectrum, May 31, 2016

Chapter 1
Networks of Social and Moral Norms in Human and Robot Agents

B.F. Malle, M. Scheutz and J.L. Austerweil

Abstract The most intriguing and ethically challenging roles of robots in society are those of collaborator and social partner. We propose that such robots must have the capacity to learn, represent, activate, and apply social and moral norms—they must have a *norm capacity*. We offer a theoretical analysis of two parallel questions: what constitutes this norm capacity in humans and how might we implement it in robots? We propose that the human norm system has four properties: flexible learning despite a general logical format, structured representations, context-sensitive activation, and continuous updating. We explore two possible models that describe how norms are cognitively represented and activated in context-specific ways and draw implications for robotic architectures that would implement either model.

Keywords Moral norms · Social norms · Norm processing · Cognitive architecture · Human-robot interaction · Robot ethics

1.1 Introduction

The design and construction of intelligent robots has seen steady growth in the past 20 years, and the integration of robots into society is, to many, imminent (Nourbakhsh 2013; Sabanović 2010). Ethical questions about such integration have recently gained

B.F. Malle (✉)
Department of Cognitive, Linguistic, and Psychological Sciences,
Brown University, Providence, USA
e-mail: bfmalle@brown.edu

M. Scheutz
Department of Computer Science, Tufts University, Medford, USA

J.L. Austerweil
Department of Psychology, University of Wisconsin, Madison, WI, USA

M.I. Aldinhas Ferreira et al. (eds.), *A World with Robots*,
Intelligent Systems, Control and Automation: Science and Engineering 84,
DOI 10.1007/978-3-319-46667-5_1

prominence. For example, academic publications on the topic of robot ethics doubled between 2005 and 2009 and doubled again since then, counting almost 200 as of the time of this conference (Malle 2015).

One set of ethical questions pertinent to robotics examines how humans should design, deploy, and treat robots (Veruggio et al. 2011); another set of questions examines what moral capacities robots themselves could have (and should have) so as to become viable participants in human society. The latter set of questions is often labeled "machine morality" (Sullins 2011) or "machine ethics" (Moor 2006), and our contribution is to this theme.

Considerations of machine morality are especially important when we assess robots in collaborative relationships with humans. A collaboration can be defined as a set of actions coordinated among two or more agents in pursuit of joint goals. An agent's pursuit of *joint* goals (rather than merely individual ones) requires several unique capacities, such as social cognition and communication. Even more fundamental, however, collaborations rely on a *norm system* that the partners share—a system that enables, facilitates, and refines the collaborative interaction (Ullmann-Margalit 1977).

As a social species, humans have become highly adept at pooling mental and physical resources to achieve goals together that they would never be able to achieve on their own. From big-game hunting to mass migration, from felling a tall tree to playing a symphony—humans work cooperatively to create common goods. But cooperative work comes with risks, because one partner might invest all the work and the other partner might reap all the benefits. Economic scholars have puzzled for a long time why such free-riding is not more common—why people cooperate much more often than they "defect," as game theorists call it, when defecting would provide the agent with larger utility.

The answer cannot be that humans are "innately" cooperative, because they are perfectly capable of defecting. The answer involves to a significant extent the power of *norms*. A working definition of a norm is the following:

> An instruction to (not) perform a specific or general class of action, whereby a sufficient number of individuals in a community (a) indeed follow this instruction and (b) demand of others in the community to follow the instruction.

Why are norms so powerful? First, they increase the predictability of other people's behavior. In a norm-guided society, any member can assume that other people will abide by norms, which greatly reduces the uncertainty over what actions they might perform. Second, norms guide a person's own action selection (especially when the optimal action is not easily determined) because norms directly tag possible actions as desirable or undesirable in the given community. Third, norms improve coordination among collaborators. That is because a collaboration involves many requests, agreements, and commitments that bind the individual to a course of action. Public promises, for example, are prototypical commitments to a norm: The declaration "I promise *X*" imposes a norm on oneself to strive toward *X*, which involves others' expectations for the person to strive toward *X*, the person's desire to

meet those expectations, and the possible sanctions other people may impose if the person fails to achieve *X*.

Norms appear to be indispensable for human social life (Hechter and Opp 2001; Ullmann-Margalit 1977). As a result, norms are likely to be indispensable for robots in human societies as well, if we expect people to perceive robots as suitable partners in effective, safe, and trusting collaborations. But what would it mean for a robot to have "norms"—whether moral norms (e.g., "do no harm") or social norms (e.g., "shake hands when meeting someone")?

Any robot involved in physical tasks will have to know a number of instrumental rules: *if an object of type F appears in area*$_1$, *move arm and grab F*. For humans, too, physical tasks require rules—actions that have high utility when certain preconditions hold. By contrast, social and moral norms are rules that are not directly dictated by a personal utility calculation (Andrighetto et al. 2010), and often they are not as action-specific as instrumental rules (e.g., "Be nice!"). Moreover, social and moral norms have other properties that make them a unique challenge for cognitive and computational examination: there seems to be an enormous number of them but they are activated extremely quickly; they are activated in highly context-specific ways but also come in bundles; they can be in conflict with one another but also can be adjusted; and they are learned fast through a variety of modalities (e.g., observation, inference, instruction).

If our goal is to build trustworthy and morally competent robot collaborators (Malle and Scheutz 2014), robots must have a computationally implemented norm system. This is because humans will demand that a robot collaborator grasps the norms of its community, and humans will withdraw their trust and cooperation if they realize that the robot does not abide by the same norms as they do.

However, we currently do not know how to incorporate sophisticated norm processing into robotic architectures. We therefore take initial steps toward a cognitive-computational model of norms by delineating core properties of the human norm system, contrasting two models of a computational norm system, and deriving implications for how robotic architectures would implement such a norm system. Ultimately, we will need to examine (1) how a cognitive system can represent and store norms, (2) how and when it activates and retrieves them, (3) how it resolves conflicts among them; (4) how it can use them in decision-making and action execution, and (5) how it can acquire them. Here we will begin to address the first two points.

1.2 Defining Norms

To begin, we introduce a general formulation of norms as consisting of three elements: a *context precondition*, a *deontic operator* ("obligatory", "forbidden", or "permitted"), and an argument that can be either an *action* or a *state*.

Specifically, let C be a context expression in a given formal language $\mathcal{L}$, and let **O**, **F**, and **P** denote the modal operators, respectively, for "obligatory", "forbidden",

and "permissable" (e.g., $\mathbf{O}\phi$ means "it is obligatory that ϕ"). Then we can provide a general schema for capturing simple norms as follows:

$$\mathcal{N} = C \rightarrow (\neg)\{\mathbf{O}, \mathbf{P}, \mathbf{F}\}\{\alpha, \sigma\} \tag{1.1}$$

The deontic operators can be analyzed cognitively as follows. To represent an action or state as *obligatory* [*forbidden*], at least three conditions must be met (Bicchieri 2006; Brennan et al. 2013)[1]:

(i) The agent represents an instruction to [not] perform a specific action or general class of action.
(ii) The agent believes that a sufficient[2] number of individuals in the reference community in fact follow the instruction.
(iii) The agent believes that a sufficient number of individuals in the reference community demands of others in the community to follow the instruction.

Conditions (ii) and (iii) are important. During the learning of a new norm and during continued application of a familiar norm, the agent must be able to update beliefs about what community members do and what they demand of one another. If the agent notices that few community members follow the instruction in question (e.g., staying within highway speed limits), then the instruction is weakened and the agent may no longer treat it as binding. And if the agent notices that few community members demand of others to follow the instruction (even though many of them still do), the instruction becomes optional and also loses its force as a norm.

These features distinguish *norms* from *goals* and *habits*, because the latter can hold even when individuals completely disregard other community members' actions or demands. Consider the action of parking one's car nose-in (in parking lots with spots marked like this: / / / /). If a majority of people perform this action but nobody actually *expects* others to do it, the action is a widely prevalent habit, not governed by a norm. And if a particular agent performs the action but is unaware that others expect him to (and they in fact do), then this agent acts to achieve a goal but does not abide by a social norm.

1.3 Properties of the Human Norm System

We propose that human norm systems have four major properties (Fig. 1.1). We first introduce each of these properties and then significantly expand on the properties of representation and activation.

[1] This is a cognitive definition of a norm, and it allows for an agent to endorse an illusory norm—when all three conditions are met but community members do not in fact follow the instruction and do not in fact expect others to follow the instruction. If we want to model and predict the agent's behavior, however, we can still consider the person to follow a perceived norm (Aarts and Dijksterhuis 2003).

[2] The threshold of sufficiency will typically be a majority but may vary by norm type and community.

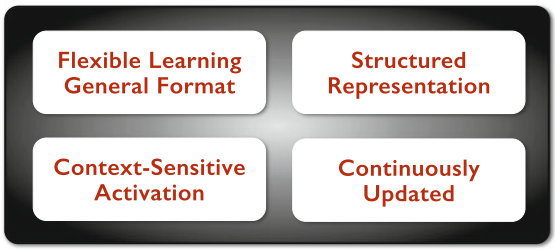


Fig. 1.1 Four cognitive properties of the human norm system

Property 1: Flexible Learning

The first property is that norm systems are learned through a variety of means (e.g., conditioning, imitation, observation, inference, and verbal instruction) but are stored in a generalized format sketched above (Eq. 1.1): as a representation of actions or states (post-conditions), given contextual preconditions. A more detailed treatment of how learning could be implemented computationally requires a better understanding of how norms are represented in the first place, and this is what we will attempt to provide shortly.

Property 2: Structured Representations

The second property is that norm systems are encoded using structured representations, systematically organized in at least three ways: *vertically* (as hierarchical layers of abstraction, ranging from action rules to general values), *horizontally* (as bundles of covarying norms tied together by the contexts in which they apply), and *temporally* (as "scripts" (Schank and Abelson 1977) that prescribe normative action sequences in a particular context, such as visiting a restaurant, greeting a friend, or boarding an airplane).

These organizing principles reflect actual features of the world. Because preconditions covary in real-world contexts (otherwise distinct contexts could not even be recognized), activated norms will also covary as bundles within contexts (horizontal organization). Likewise, because the human action planning and execution system is organized hierarchically and temporally, norms that guide such action will incorporate this organization as well.

The structured organization of norms is likely to have far superior processing characteristics than the simplest alternative—(long) lists of singlet norms. That is because norms can be thought of as nodes in a memory network, and we know that structured organization of memory representations has significant advantages in memory accuracy, efficiency, and speed of retrieval (Bower 1970).

Property 3: Context-Sensitive, Bundled Activation

A third property, we suggest, is that specific contexts rapidly activate norms as connected bundles. There is evidence that norms are indeed activated in highly context-specific ways (Harvey and Enzle 1981; Aarts and Dijksterhuis 2003; Cialdini et al. 1991) and that norm violations are detected very quickly (Van Berkum et al. 2009).

These characteristics are responses to a world in which a large number of norms exist but only a small subset is relevant in any given context. The norm system therefore must be both comprehensive in its representational capacity and selective in its activation patterns. These demands pose numerous challenges for the computational implementation of a norm network, so we will dedicate much of our subsequent analysis to these challenges.

Property 4: Continuous Updating

The fourth property of the human norms system is that the context-sensitive norm networks are continuously updated—for example, when a new norm is learned or a new context is added as a precondition to a previously learned norm. This makes the norm system highly flexible when people encounter "mixed" contexts, mixed roles, or enter unfamiliar communities. It also allows for rapid societal change—whether due to natural events (e.g., climate change), technological innovation (e.g., the internet), or collective preferences (e.g., gay marriage).

Cognitively speaking, when a context is added as an additional precondition for a given norm, the likelihoods of co-activation (bundling) among norms will change because these likelihoods are a direct function of the number of preconditions shared between norms. How quickly the likelihoods change will depend on general principles of the norm network. For example, updating will be frequent if co-activation of two norms instantly forms a direct connection between them. Likewise, updating will be frequent if equivalence between contexts is loose (i.e., features that define contexts are correlated both within and between contexts, rather than figuring as necessary and sufficient conditions).

We now turn to the central portion of our chapter: an analysis of how norms, defined as context-specific instructions, can be activated in bundles tailored to their particular contexts.

1.4 Challenges of Context-Sensitive, Bundled Norm Activation

We have argued that norms are activated in *specific contexts* and as *connected bundles*. How can we account for these characteristics? We first outline the logical format of these bundles and then consider potential computational models of how they are represented and activated.

1.4.1 Logical Format

Expressed in the logical format of Eq. 1.1, each norm has a set of preconditions C that correspond to contexts in which the norm applies (e.g., $C \rightarrow \mathbf{F}\phi$) or in which the norm is specifically suspended ($C \rightarrow \neg\, \mathbf{F}\phi$). When a given situation Σ meets the

contextual preconditions C of a given norm, the norm will be quickly activated. The critical open question here is what "meeting the contextual preconditions" means.

Let f_Σ be the set of features present in a given situation Σ and let f_C be the features that constitute the preconditions C for a given norm. We hypothesize that the degree of activation of the norm in a given situation Σ will be a function of the number of features shared between the situation and the preconditions of the norm (e.g., $|f_\Sigma \cap f_C|$, where $|\cdot|$ is the cardinality of a set, possibly weighted and scaled by factors depending on the contextual features and the norm). If this hypothesis is correct, then *all* norms that have any $f \in \Sigma$ in their set of preconditions ($f \in C$) will be activated to *some degree*. We will call these co-activated norms in a given situation Σ "norm bundles."

Note that for two norms $\mathcal{N}_1$ and $\mathcal{N}_2$ in a norm bundle it could be the very same contextual property $f_i \in \Sigma$ that is in both of their norm preconditions ($f_i \in C_1$ and $f_i \in C_2$). Alternatively, different features in the situation ($f_1, f_2 \in \Sigma$) could activate different norms ($f_1 \in C_1$ and $f_2 \in C_2$, but $f_2 \notin C_1$ and $f_1 \notin C_2$). Hence,

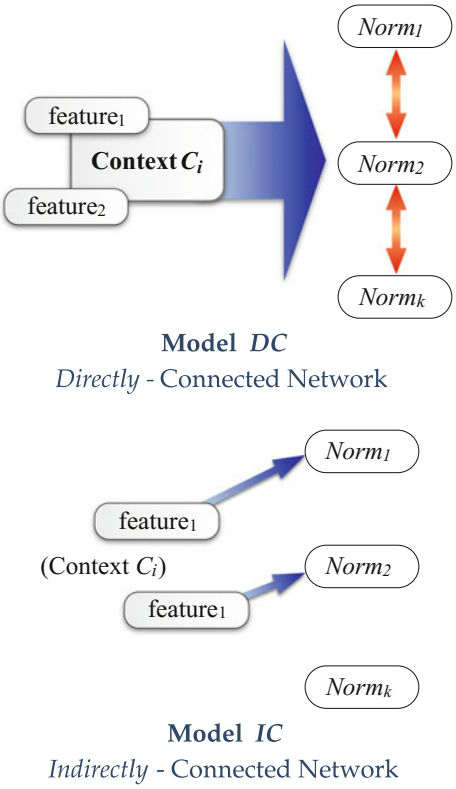


Fig. 1.2 Two models of how contexts can activate "bundles" of norms. Under the first, but not the second, model, $Norm_k$ would be activated

norm bundles do not necessarily have to share any particular situational features, even when their constituent norms are co-activated, as long as there are reliable co-variations of situational features. From a computational perspective the question then arises exactly how these co-variations are represented; that is, whether norms in norm bundles are represented in the human cognitive architecture as connected *directly* with one another or connected only *indirectly*, via the shared preconditions between the norms and the situational features that trigger them. Hence, there are at least two different models of how such covariation among norms in a norm bundle can come about—models that specify in what way norms are part of a "bundle" (see Fig. 1.2).

1.4.2 Two Models of Norm Covariation

In a *directly-connected* network (Model *DC* in Fig. 1.2), norms and their co-activation are represented as nodes and edges in a mathematical network, where each edge is given a weight indicating the strength of association between the nodes (Harvey and Enzle 1981), possibly built up through learning and repeated co-activation. A given context (constituted by a fuzzy set of features) activates a particular norm network in part because the context activates some norms and these norms activate other, connected norms.

Alternatively, in an *indirectly-connected* network (Model *IC* in Fig. 1.2), specific features (e.g., objects in a scene) independently activate specific norms, and sets of norms covary as bundles solely because the features that activate them typically co-vary within contexts, not because of direct connections among the norms themselves. In this more minimalist network, no additional concept of a "context" (above and beyond an extensional class of features) needs to be postulated. The "affordances" of objects and properties in scenes suffice to activate the right kinds of norms.

For example, holding the fork a certain way and holding the knife a certain way while eating at the table may be a connected pair of norms that is activated as a bundle by the sight of a set table; alternatively, the fork may activate *its* norm of use and the knife may activate *its* norm of use, and the two norms are co-activated merely because, in the real world, knives and forks are typically co-present.

1.4.3 Different Empirical Predictions

Although both models account for "norm bundling," the two models make different predictions about norm activation patterns in unusual situations. Consider a situation Σ (e.g., eating at a fine-dining restaurant) that is normally constituted by a sufficient subset from the set of features f_1–f_6 and, if recognized as a particular context C, reliably activates the bundle of norms $\mathcal{N}_1$–$\mathcal{N}_4$, which all have C as their precondition. Now suppose that the perceptual input is impoverished (e.g., bad light-

ing or intense noise), making only features f_1–f_3 available in this particular case. According to the *directly-connected* model, such an impoverished scene would still be likely to activate the whole bundle of norms, because even a few directly activated norms would themselves activate other norms with which they normally covary. By contrast, according to the *indirectly-connected* model, norms are activated only by specific features (e.g., objects) in a scene, and therefore the impoverished situation would elicit "incomplete" norm bundles—only those that are individually activated by features f_1–f_3.

Likewise, the models make different predictions when a foreign object is embedded into a scene (e.g., a baseball in a fine-dining restaurant). According to the *DC* model, a foreign object would have little effect on the activated norms, because once an overall context triggers its bundle of norms, any effect of specific (additional) features would be drowned out (or at least mitigated). Not so for the *IC* model, according to which norms are activated individually by specific features (e.g., objects) in the scene. The baseball in the restaurant would have a marked effect on the set of activated norms, because people cannot help but bring to mind whatever one may (or may not) do with a baseball, even in a fine-dining restaurant.

1.4.4 Implications for Cognitive Robotic Architectures

Implementations of the *DC* model in cognitive robotic architectures could be analogous to networks of spreading activation (e.g., in the spirit of the declarative memory in ACT-R) where a given context (constituted by a sufficient subset of features) will spread activation to the norms that have this context as a precondition. As mentioned, the norms in a given bundle need not have a single precondition that is shared among all of them—as long as some of the norms share some preconditions with other norms and some subset of these partially shared preconditions are present, the bundle will be activated through spreading activation. The main advantage of directly-connected norm bundles is that partial matches or inaccurate perceptions may still be sufficient to activate all norms in a bundle. This is because the direct connections among norms within a bundle will spread activation to each other, so as long as some of the norms are immediately activated (e.g., through perceptions, inferences, etc.), the other ones will eventually become activated as well. The main disadvantage of directly-connected norm bundles is that some norms might become inappropriately activated (i.e., without there being a contextual feature to which the norm applies), simply because direct linkages can drag one norm along with another.

Implementations of the *IC* model, on the other hand, do not require representational mechanisms such as spreading activation, as all norms in a bundle will be solely activated by the situational features that match their context preconditions. Hence, the main advantage of indirectly connected norm bundles is that the norms are activated in close correspondence to situations and their recognizable or inferable features. Such a network need not engage in inferences about "contexts" as separate constructs, because contexts are merely extensional classes of features. Of course, if

features are highly correlated, such extensional classes could be learned as higher-level categories, but they do not have to be separately represented each time a norm is activated. The main disadvantage of indirectly-connected norm bundles is that acute and fast perceptual processes are required that recognize all relevant objects and properties in the environment so as to activate their corresponding norms (e.g., permissible ways of handling a fork, a knife, a spoon, a napkin, ...).

Critically, however, both models require ways to arbitrate among activated norms that have mutually contradictory implications. For example, norm $\mathcal{N}_1$ might impose an obligation to do A while $\mathcal{N}_2$ might impose an obligation to do B, yet either doing A and B is not possible at the same time, or doing one of them will undo prerequisites of the other in a way that the other action can no longer be performed.

Deciding between the two models will also influence the general logical form of norms. If there are direct connections between, say $\mathcal{N}_1$ and $\mathcal{N}_2$ (above and beyond shared preconditions, i.e., context features), how are these connections represented? Are they continuous and/or probabilistic? And what implications does such a representation have for logical reasoning on deontic operators? If, on the other hand, there are no connections among norms themselves, can we completely characterize norm networks as arrays of context features that do or do not activate specific norms? We next explore these possibilities in more detail.

1.4.5 What Would Constitute Norm Connections?

Figures 1.3 and 1.4 illustrate how quantitative predictions for norm co-activation strength can be derived from each model. Figure 1.3 shows a hypothetical norm system represented in a table where rows index features f_1–f_6 that constitute contexts C_1–C_4 and columns index norms that can be activated by these features. A cell is 1 if the corresponding norm is activated in the presence of the feature (and 0 otherwise, but left empty in the table for better readability).

According to the *indirectly*-connected model, the strength of co-activation of norm $\mathcal{N}_i$ with $\mathcal{N}_j$, the formula $r_f(\mathcal{N}_i, \mathcal{N}_j)$, can be written as:

$$r_f(\mathcal{N}_i, \mathcal{N}_j) = \frac{\sum_C \sum_f I(f \in \mathcal{N}_i \wedge \mathcal{N}_j)}{\sum_C \sum_f I(f \in \mathcal{N}_i)}, \tag{1.2}$$

where $I(\cdot)$ is the identity function that returns 1 when its argument is true and 0 otherwise. According to Eq. 1.2, the strength of co-activation of $\mathcal{N}_i$ with $\mathcal{N}_j$ is the number of features (repeating features over contexts) they have in common, normalized by the number of features that are preconditions for $\mathcal{N}_i$ (again repeating features over contexts). For example, focusing on context C_2, feature f_3 co-activates norms $\mathcal{N}_1, \mathcal{N}_3$, and $\mathcal{N}_4$; feature f_4 co-activates $\mathcal{N}_3$ and $\mathcal{N}_4$; and feature f_5 co-activates $\mathcal{N}_1$ and $\mathcal{N}_4$. Features can reappear across contexts, and this is illustrated above by the fact that f_3 and f_5 also help constitute context C_4. All these co-activation patterns

of norms, triggered by features, lead to the feature-level co-activation matrix on the top table of Fig. 1.4.

According to the *directly*-connected model, what counts are not feature-level co-activations but context-level co-activations. Contexts, latent factors inferred from slightly varying sets of features, activate their norms *as a set*, with some norms activated by already activated other norms, not by features. Thus, according to the *directly*-connected model, the strength of co-activation between norm $\mathcal{N}_i$ and $\mathcal{N}_j$, the formula $r_c(\mathcal{N}_i, \mathcal{N}_j)$, is:

$$r_c(\mathcal{N}_i, \mathcal{N}_j) = \frac{\sum_C I(\exists f \in C : f \in \mathcal{N}_i \wedge f \in \mathcal{N}_j)}{\sum_C I(\exists f \in C : f \in \mathcal{N}_i)}. \tag{1.3}$$

According to Eq. 1.3, strength of co-activation is the ratio of the number of contexts where both $\mathcal{N}_i$ and $\mathcal{N}_j$ are applicable to the number of contexts where $\mathcal{N}_i$ is applicable. For example, f_3 and f_4 would be taken as sufficient evidence for the presence of C_2, and C_2 would activate, as a set, $\mathcal{N}_1$, $\mathcal{N}_3$, and $\mathcal{N}_4$. No matter which features in a scene allow a given context to be inferred, all of its norms (the norms that have that context as a precondition) are activated, and co-activation among these norms leads, over time, to norm interconnections. Those are represented as context-level connection strengths (again normed against number of norm occurrences) in the bottom table of Fig. 1.4.

We see that the two matrices are quite different, so they should in principle be empirically distinguishable. Mere feature-caused co-activation predicts far smaller co-occurrence frequencies than context-caused co-activation with subsequent connection formation. If we can measure such norm co-activations (and we are currently developing a paradigm to do so), we have yet another way of arbitrating between the two models, which would teach us about the underlying principles of human norm

Context	feature	$\mathcal{N}_1$	$\mathcal{N}_2$	$\mathcal{N}_3$	$\mathcal{N}_4$
C_1	f_1		1	1	
	f_2		1		
C_2	f_3	1		1	1
	f_4			1	1
	f_5	1			1
C_3	f_6	1	1		
C_4	f_3	1		1	1
	f_2		1		
	f_6	1	1		
	f_5	1			1

Fig. 1.3 Contexts (C_1–C_4) and their features (f_1–f_6) that activate specific norms $\mathcal{N}_1$–$\mathcal{N}_4$. Cells with unique colors indicate co-activation of two or more norms by a particular feature

Feature-level co-activation

		$\mathcal{N}_1$	$\mathcal{N}_2$	$\mathcal{N}_3$	$\mathcal{N}_4$
$\mathcal{N}_1$	with		2/6	2/6	4/6
$\mathcal{N}_2$	with	2/5		1/5	0/5
$\mathcal{N}_3$	with	2/4	1/4		3/4
$\mathcal{N}_4$	with	4/5	0/5	3/5	

Context-level co-activation

		$\mathcal{N}_1$	$\mathcal{N}_2$	$\mathcal{N}_3$	$\mathcal{N}_4$
$\mathcal{N}_1$	with		2/3	2/3	2/3
$\mathcal{N}_2$	with	2/3		2/3	1/3
$\mathcal{N}_3$	with	2/3	2/3		2/3
$\mathcal{N}_4$	with	2/2	1/2	2/2	

Fig. 1.4 Computation of norm co-activation at the level of features (*top* table) and at the level of contexts (*bottom* table)

networks and provide benchmarks for corresponding norm networks in robotic architectures.

We should add that the two models *DC* and *IC* also make different predictions about the process of norm updating (Property 4 mentioned earlier). When a context is added as an additional precondition for a given norm, the *DC* would predict that this norm soon picks up new connections with other norms, because the co-activation (bundling) likelihoods among norms are a direct function of the number of shared preconditions between norms. According to the *IC* model, by contrast, the norm co-activation pattern changes more slowly, and only to the extent that the pattern of feature co-occurrences changes.

Clearly, a number of hybrid models could be constructed as well. For example, one model could allow norm-to-norm interconnections without postulating contexts as latent factors inferred from features. In this case, features directly cause norm co-activation *and thereby* cause formation of real norm interconnections, so norms could also be activating each other (e.g., $f_4 \rightarrow \mathcal{N}_4 \rightarrow \mathcal{N}_3$). The problem that arises for a network with these characteristics is that norms could activate other norms that are *not* appropriate for a given context. Consider C_1 in the example norm system of Fig. 1.3. If $f_1 \rightarrow \mathcal{N}_3$ and, because of the strong interconnection $r_f(\mathcal{N}_3, \mathcal{N}_4)$, also $f_1 \rightarrow \mathcal{N}_4$, then the norm $\mathcal{N}_4$ is activated in C_1 even though, by assumption for this example network, it shouldn't be active in this context. Thus, the model may have to incorporate inhibitory connections in addition to excitatory connections—which would then lead to interesting new predictions.

This highlights the general question of how the human norm network cognitively instantiates an intuitive requirement: that contexts reliably activate the "right" bundle of norms, not just some bundle of previously co-occurring norms. Achieving this reliability is made difficult by the fact that contexts are likely to show fluctuation in the specific set of features that instantiate a context in any particular case. A *DC* net-

work relies on inferred context categories built right into the cognitive system, which creates robust invariance across feature fluctuations (because the learned norm-to-norm interconnections maintain the identity of context categories). An *IC* network would be far more sensitive to feature fluctuations. Every time a new feature combination emerges, it triggers a slightly different set of norms. So equivalence classes for what is the "same context" would be difficult to form. But because the *IC* model does not rely on abstract context representations and instead responds to natural, complex feature intercorrelations (that may, in reality, constitute true contexts), the reliability and invariance of the norm network is a direct function of the reliability and invariance of the world itself—the more the world fluctuates, the more an *IC* network offers finely adjusted sets of activated norms.

1.4.6 In Dictu *Norm Activation*

So far we have analyzed norm activation in situ—that is, in real-world situations that offer a rich array of features, which can constitute contexts. But norm activation (and indeed, norm learning) often occurs in dictu, when one person tells another person to (not) act in a certain way "in church" or "when adults are around" or "when somebody just experienced a loss". What would the *indirectly-connected model* say about such situations? Where are the specific features that would trigger the specific norms? Is this not a case in which contexts are like latent factors that directly trigger a bundle of norms that have become interconnected?

This situation does not actually cause a problem for the *IC model*. A minimalist model about norm interconnections does not have to be minimalist about concept-feature and feature-feature interconnections. It would be strange to deny, in light of the semantic network and category literature, that concepts such as "in church" could not activate a large number of features that then directly activate norms. The idea that context categories directly activate norms is in fact less plausible because the fuzziness of categories such as "in church" (in the physical building? in a cathedral? during mass?) doesn't easily select for specific bundles of norms. The addressee would have to disambiguate the vague category (either in their own mind or by asking questions) and thereby "fix" the relevant features, which in turn would activate relevant norms.

1.4.7 Context and Structured Organization

We have illustrated how context interacts with the horizontal structural organization of norms—their direct connections or indirect co-activation patterns. Context can also exert a powerful influence on norm activation by means of vertical (hierarchical) structures in the norm system. When planning to go to a business meeting, for example, abstract norms such as "be respectful" might be activated merely by think-

ing about the meeting in advance. "Be respectful" by itself does not have specific action instructions, but when a moment arises in which a business partner says something obviously incorrect, the norm may translate down the abstraction hierarchy into a concrete instruction to remain quiet or to be expressly hesitant in one's correction.

Context categories can also exert a powerful influence on norm activation by means of temporal structures. Driving up to the restaurant and reading the "Valet Parking" sign triggers a normative sequence of actions (parking the car at the sign, greeting the valet, passing the key, accepting a number tag, etc.). The activated norm may "reel off" a series of sequential instructions that are associated with one another, not necessarily as *norm* interconnections but as well-practiced *action* interconnections.

1.5 Summary and Conclusion

For a robot to become ethical it will need to have a *norm capacity*—a capacity to learn, represent, activate, and apply a large number of norms that people expect one another to obey and, in all likelihood, will expect robots to obey. To build such a norm capacity we will need to make critical decisions about how a norm system is organized and implemented in the robot's cognitive architecture. We have focused on the contrast between two models of how such a norm system might be organized—as *directly* or *indirectly* connected networks—and illustrated some of the questions that this contrast raises. At the same time, we have set aside countless other questions. For example, in designing a robot's norm network, how would the specific norms that apply within a community, as well as their triggering contexts, be identified? How would a computational norm network handle norm conflict—that is, cases in which features in a given situation activate norms with contradictory action instructions or incompatible state goals. And exactly how can a system expand and refine its norm network without suffering from serious interference among its norms? Despite the many unanswered questions, we hope that delineating key properties of the human norm system and beginning to analyze logical and computational characteristics of this system will prove fruitful in the endeavor to make robots socially and morally acceptable participants in society.

Acknowledgements This work was in part funded by a grant from the Office of Naval Research (ONR), No. N00014-14-1-0144, and a grant from the Defense Advanced Research Projects Agency (DARPA), DARPA SIMPLEX No. 14-46-FP-097. The opinions expressed here are our own and do not necessarily reflect the views of ONR or DARPA.

References

Aarts H, Dijksterhuis A (2003) The silence of the library: environment, situational norm, and social behavior. J Pers Soc Psychol 84(1):18–28

Andrighetto G, Villatoro D, Conte R (2010) Norm internalization in artificial societies. AI Commun 23(4):325–339
Bicchieri C (2006) The grammar of society: the nature and dynamics of social norms. Cambridge University Press, New York
Bower GH (1970) Organizational factors in memory. Cogn Psychol 1(1):18–46. doi:10.1016/0010-0285(70)90003-4
Brennan G, Eriksson L, Goodin RE, Southwood N (2013) Explaining norms. Oxford University Press, New York
Cialdini RB, Kallgren CA, Reno RR (1991) A focus theory of normative conduct: a theoretical refinement and reevaluation of the role of norms in human behavior. In: Zanna MP (ed) Advances in experimental social psychology, vol 24. Academic Press, San Diego, CA, pp 201–234
Harvey MD, Enzle ME (1981) A cognitive model of social norms for understanding the transgression helping effect. J Pers Soc Psychol 41(5):866–875. doi:10.1037/0022-3514.41.5.866
Hechter M, Opp KD (eds) (2001) Social norms. Russell Sage Foundation, New York
Malle BF (2015) Integrating robot ethics and machine morality: the study and design of moral competence in robots. Ethics Inf Technol [online first]. doi:10.1007/s10676-015-9367-8
Malle BF, Scheutz M (2014) Moral competence in social robots. In: IEEE international symposium on ethics in engineering, science, and technology. IEEE, Chicago, IL, pp 30–35
Moor JH (2006) The nature, importance, and difficulty of machine ethics. IEEE Intell Syst 21(4):18–21. doi:10.1109/MIS.2006.80
Nourbakhsh IR (2013) Robot futures. MIT Press, Cambridge
Schank RC, Abelson RP (1977) Scripts, plans, goals, and understanding. Erlbaum, Hillsdale
Sullins JP (2011) Introduction: open questions in roboethics. Philoso Technol 24(3):233. doi:10.1007/s13347-011-0043-6
Šabanović S (2010) Robots in society, society in robots. Int J Social Robot 2(4):439–450. doi:10.1007/s12369-010-0066-7
Ullmann-Margalit E (1977) The emergence of norms. Clarendon library of logic and philosophy. Clarendon Press, Oxford
Van Berkum JJA, Holleman B, Nieuwland M, Otten M, Murre J (2009) Right or wrong? the brains fast response to morally objectionable statements. Psychol Sci 20(9):1092–1099. doi:10.1111/j.1467-9280.2009.02411.x
Veruggio G, Solis J, Van der Loos M (2011) Roboethics: ethics applied to robotics. IEEE Robot Autom Mag 18(1):21–22. doi:10.1109/MRA.2010.940149

Chapter 2
Perceived Autonomy of Robots: Effects of Appearance and Context

Maaike Harbers, Marieke M.M. Peeters and Mark A. Neerincx

Abstract Due to advances in technology, the world around us contains an increasing number of robots, virtual agents, and other intelligent systems. These systems all have a certain degree of autonomy. For the people who interact with an intelligent system it is important to obtain a good understanding of its degree of autonomy: what tasks can the system perform autonomously and to what extent? In this paper we therefore present a study on how a system's characteristics affect people's perception of its autonomy. This was investigated by asking fire-fighters to rate the autonomy of a number of search and rescue robots in different shapes and situations. In this paper, we identify the following seven aspects of perceived autonomy: time interval of interaction, obedience, informativeness, task complexity, task implication, physical appearance, and physical distance to human operator. The study showed that increased disobedience, task complexity and physical distance of a robot can increase perceived autonomy.

Keywords Autonomy · Robots · Intelligent agents · Intelligent systems · Robot design · User expectations · Human-robot interaction · Perceived autonomy

M. Harbers (✉) · M.M.M. Peeters (✉)
Delft University of Technology, Delft, The Netherlands
e-mail: M.Harbers@tudelft.nl

M.M.M. Peeters
e-mail: M.M.M.Peeters@tudelft.nl

M.A. Neerincx
TNO Human Factors and Delft University of Technology, Delft, The Netherlands
e-mail: mark.neerincx@tno.nl

M.I. Aldinhas Ferreira et al. (eds.), *A World with Robots*,
Intelligent Systems, Control and Automation: Science and Engineering 84,
DOI 10.1007/978-3-319-46667-5_2

2.1 Introduction

Our environment contains an ever increasing number of robots in all sorts and forms. Examples of contemporary robots include (self-driving) cars, smart ambient home systems, vacuum cleaners, lawn mowers, virtual assistants, stuffed animals, search and rescue robots, and robots consisting of a set of stationary mechanical arms, such as laparoscopic surgery robots and assembly line robots. As a result of this gradual move of robots from contained environments to social, human environments, the group of people interacting with robots—especially service robots—is no longer confined to computer experts alone, but also includes non-expert users.

Robots have different degrees of autonomy, i.e., robots have different capabilities and differ in the extent to which they can perform tasks independently. People interacting with a robot that is new to them often estimate the robot's capabilities based on its observable characteristics (Blow et al. 2006; Kaplan 2005). Users can be disappointed in a robot when their estimations of the robot's capabilities do not match its actual ones (Kaplan 2005). Underestimation of a robot's capabilities can lead to the robot not being exploited to its fullest, and overestimation may result in the robot being deployed for tasks exceeding its capabilities (Hancock et al. 2011). Both of the above are undesirable, yet laymen are not always capable of making accurate estimations of robot capabilities.

Most people's conception of what a robot is appears to be largely based on the way robots are depicted in fiction (Broadbent et al. 2010; Lohse et al. 2008). The term 'robot' was first coined to denote fictional automata in a 1925 play called 'R.U.R.—Rossum's Universal Robots' by Čapek (1925). Since then, robots have featured in movies, books and games, and they have been sold as toys (Telotte 1995).

There are large differences in the type of roles assigned to robots in fiction. In some stories they are depicted destroying the world or seeking world-domination to rule over humans. Other stories feature robots as human-like compassionate entities, such as personal assistants or even like-minded friends. Yet, robots in fiction are largely presented as independent, autonomous actors that have a 'mind of their own', with a humanoid or anthropomorphic appearance. Thus, laymen often conceive a robot as an acting and thinking anthropomorphic entity, confined to a body resembling that of a human (Duffy 2003).

There are some notable differences between the way robots are presented in fiction and the way they actually occur in real life. In the field of robotics, robots are usually considered as computer-controlled machines that can perceive and manipulate their physical environment (Spong et al. 2006). In this sense of the concept, most smart devices, e.g. smart televisions or phones, are considered robots. In contrast to fictional robots, these robots highly differ in what they are able to do and the extent to which they can perform tasks autonomously. Some robots are able to perform well-constrained tasks—such as surgery, driving on a highway, or vacuum cleaning—completely autonomous. Other robots are tele-operated by humans, and are not autonomous at all. Though some robots in real life have impressive human-

like appearances, none of them has a level of autonomy that comes close to that of humans.

Research has shown that the design (i.e., the 'look and feel') of a robot influences a user's expectation of the robot's physical and behavioral capabilities (Blow et al. 2006; Kaplan 2005). When interacting with robots that look like a human or animal, laymen tend to expect more complex behavior—or have a harder time estimating the complexity thereof—than they would in the case of a robot that more or less resembles existing appliances, such as a phone or a lawn mower (Dautenhahn 2002). It is not clear yet, how other observable robot characteristics contribute to the perception of a robot's autonomy, i.e., what it can and cannot do independently. This paper therefore presents a study that investigates how robot features influence a user's expectation of a robot's autonomy. The insights obtained in this study can be used to design robots in such a way that the user's estimations of a robot's capabilities match its actual ones.

The outline of this paper is as follows. In Sect. 2.2, we discuss several robot features that may affect a robot's perceived autonomy. In Sects. 2.3 and 2.4, we describe the methods and results of our study with fire-fighters, respectively. In Sect. 2.5 we provide a discussion and a conclusion.

2.2 Dimensions of Perceived Robot Autonomy

The Oxford dictionary defines autonomy as 'the right or condition of self-government' (Oxford Dictionaries 2015). The term is often used, both in technical and everyday language, to describe robot behavior. Apparently, people have a notion of what 'robot autonomy' means, and are able to perceive and express the extent to which they think a robot is autonomous.

Upon closer inspection, however, autonomy is a complex term. There are several misconceptions associated to the term, in particular when applied to robots (Bradshaw et al. 2013). First, autonomy is not an all-or-nothing feature that a robot either has or does not have. The concept 'levels of autonomy' is often used to describe technology that is partially autonomous. However, there is no agreement among scholars on what types of behavior should be classified as being more autonomous or less autonomous (Beer et al. 2014). Second, autonomy is a multi-dimensional concept. Johnson and colleagues, for instance, pointed out that for an entity to act autonomously, it must both be able and allowed to perform some action (Bradshaw et al. 2013; Johnson et al. 2011, 2014). Third, a robot's ability to perform an action is task-specific and context-specific, making it impossible to compare different types of behavior along a single scale (Beer et al. 2014; Johnson et al. 2011; Murphy and Shields 2012). As long as there is no entity that can perform all possible tasks in all possible circumstances, full autonomy does not exist (Bradshaw et al. 2013).

The misconceptions pointed out above seem to implicate that there are multiple factors that determine a robot's level of autonomy. In our efforts to understand how humans form an idea of a robot's autonomy on their first encounter, we therefore

distinguish seven factors that potentially explain the perceived autonomy of a robot. These seven factors are partly inspired by the following three dimensions of autonomy introduced by Scharre and Horowitz (2015) to clarify the concept of autonomy: (1) the human-machine command-and-control relationship, (2) the complexity of the machine, and (3) the type of decision being automated. We adopt the last two dimensions without alterations as factors that may explain and predict perceived autonomy, and we will discuss them in more detail later on in this section. The first dimension, we adopt, albeit with some considerable alterations as we show in the following.

Along the first dimension, Scharre and Horowitz distinguish a human in-the-loop, on-the-loop or out-of-the-loop human-machine relationship. A human in-the-loop relationship means that the robot needs human input at regular time intervals in order to proceed its actions. A human on-the-loop relationship means that the robot acts by itself, but that the human can intervene in the robot's actions at any time, e.g. veto a planned action or change the robot's goals. In an out-of-the-loop relation, the robot acts independently for certain periods of time, and in these time spans, the human has no influence on the robot's actions.

We believe that this distinction of three human-machine relationships is useful, yet insufficient to express the full range of relationship possibilities. For instance, it is not possible to express the time periods during which the human is not able to intervene in the robot's behavior in an out-of-the-loop relationship. This is important, because, for example, as these intervals become smaller, the difference between in-the-loop and out-of-the-loop become less clear. In this paper we therefore propose to unravel the human-machine command-and-control relationship into the following three factors: the time interval of interaction, the obedience of the robot and the informativeness of the robot. We will later explain these factors in more detail. We believe that these three factors allow for a more accurate expression of different human-machine relationships.

The perceived autonomy factors described so far concern a robot's actual capabilities and autonomy. The focus of our study, however, is perceived autonomy. We therefore introduce two more factors that may explain a user's estimation of a robot's autonomy: the robot's physical appearance and the physical distance between a robot and its operator.

In total, we now mentioned seven factors that may explain and predict perceived autonomy: time interval of interaction, obedience, informativeness, task complexity, task implications, physical appearance and physical distance. In the remainder of this section, we will discuss for each factor why we believe that it may affect perceived autonomy, and how we expect it to affect perceived autonomy.

2.2.1 Time Interval of Interaction

Time interval refers to the time during which a robot can act independently, i.e., without human interference. This time interval can be determined by assessing a

robot's neglect tolerance. Neglect tolerance is a measure of how the robot's current task effectiveness declines over time when the robot is neglected by the user. Several scholars have pointed out that neglect tolerance is an important metric in measuring the autonomy of a robot with respect to some task (Beer et al. 2014; Goodrich and Olsen 2003; Olsen and Goodrich 2003). Robots with a higher neglect tolerance generally need to be more autonomous in order to remain effective. We thus expect that robots acting independently for larger time intervals are perceived as more autonomous.

2.2.2 Obedience

All robots receive human input. To the very least, robots are switched on and off by a human. Most often, however, robots receive human instructions in between. Assuming that the robot understands the instructions, it may or may not choose to follow them, i.e. be obedient or disobedient. We generally want robots to be obedient (Bradshaw et al. 2013; Johnson et al. 2011, 2014). But there may be some cases where we want them to be disobedient. Take for instance a robot receiving conflicting instructions: it is asked to perform an action that threatens a human's safety. In such situations, we may prefer robots that make their own choices and are not strictly obedient. Such robots require the capability to reason autonomously about the situations they are confronted with, rather than reactively following all instructions they receive. For this reason, we expect that a robot that is occasionally disobedient for a good reason will be perceived as more autonomous.

2.2.3 Informativeness

Informativeness refers to the extent to which the robot informs humans about its capabilities, goals, plans, and current status. This property is sometimes referred to as observability (Johnson et al. 2014). We prefer the term informativeness, however, because not all computer or robot output is equally relevant and understandable to humans (Harbers et al. 2010), and the term informativeness implies that the provided data are not only observable, but also understandable and informative.

Endowing systems with the capability to provide information and explanations to its user has been shown to improve their usability (Ye and Johnson 1995). The provided information not only improves users' acceptance and understanding of a system's decisions and recommendations, it also increases their confidence in the robot's decision-making capabilities. Providing information thus increases users' expectations of a system's capabilities, which are closely related to the system's degree of autonomy. We therefore expect that informative robots are perceived as more autonomous.

2.2.4 Task Complexity

As mentioned above, Scharre and Horowitz pointed out that autonomy can refer to the complexity of a system (Scharre and Horowitz 2015). According to our notion of autonomy, a system is autonomous when it acts independently. However, systems that perform simple tasks independently are usually called automatic or automated, rather than autonomous. The term 'automatic' is often used for systems that perform simple tasks, e.g. a mechanical thermostat or an industrial robot. 'Automated' usually refers to rule-based systems such as a programmable thermostat or a diagnose support system. The term 'autonomy' is typically reserved for systems that execute some kind of self-direction, self-learning or emergent behavior. We therefore expect that robots that perform tasks of higher complexity are perceived as more autonomous.

2.2.5 Task Implications

Different tasks and decisions have different levels of risk and implications (Scharre and Horowitz 2015). A toaster and a land mine both perform tasks that are relatively simple—they both have to "go of" at some point. However, the consequences of the land mine's actions are much bigger than those of the toaster. Tasks with larger implications are generally performed by people with higher levels of responsibility and they require a larger range of competencies. We therefore expect that robots that perform tasks with larger implications are perceived as more autonomous.

2.2.6 Physical Appearance

There is a lot of evidence showing that the physical appearance of a robot influences people's perception and expectations of that robot. Lohse et al., for instance, found that appearance plays a crucial role in the perception of a robot and that it determines what types of tasks and activities are regarded as most suitable for the robot (Lohse et al. 2008). In their study, animal-like robots were merely seen as toys, whereas humanoid robots were perceived as more serious in nature. Results from a study by Goetz et al. showed that users preferred robots whose looks and behavior matched the users' expectations (Goetz et al. 2003). In addition, users would sooner comply with the robot's instructions. Walters et al. found that participants tended to prefer robots with more human-like appearances and attributes (Walters et al. 2008). Based on the above, we expect that a more human-like appearance is perceived as more autonomous.

Besides the effects of different robot appearances on perception, the difference in effect of physical versus virtual robots has been studied. Mirelman et al. found that training with an actual robot was more successful than training with a virtual robot

(Mirelman et al. 2009). Research showed that participants empathized more with a physically embodied robot than with a robot without a physical body (Kwak et al. 2013; Looije et al. 2012). Embodied robots thus seem to have stronger effects on people than disembodied ones. We therefore expect that a physical robot is perceived as more autonomous than a virtual robot.

2.2.7 Physical Distance

Research on the effect of geographic distance on human-human collaboration shows that people initially cooperate less when they believe that their collaborator is farther away (Bradner and Mark 2002). The same study also showed that people are more likely to deceive, and are less persuaded by collaborators at a larger distance from them. As physical distance affects the way humans perceive other humans, we expect that it will also affect the way they perceive robots. Furthermore, robots that are situated farther away from their operator have less access to the operator's help, they thus seem to require a higher level of independence. We therefore expect that robots situated farther away from their operator are perceived as more autonomous.

2.3 Methods

We performed a study in the domain of search and rescue robots to measure the extent to which the factors identified in the previous section contribute to perceived robot autonomy. The setup of our study was as follows.

Participants. 18 voluntary or professional firefighters participated in this study. Their work experience ranged between 3 and 30 years. Three of them had previous experience with search and rescue robots.

Measures. For our study we developed a questionnaire on perceived autonomy (see http://ii.tudelft.nl/perceived-autonomy). It opens by asking for a definition of the term autonomy. On the next page of the questionnaire, it is stated that the term 'autonomy' will be used to mean 'acting independently' throughout the rest of the questionnaire. Subsequently, a picture depicting a robot is displayed and participants are asked to indicate how autonomous they consider this system to be (Fig. 2.1). The participants are instructed to use their intuition.

The first picture displays the robot under 'normal circumstances' and serves as a baseline measure. This baseline question is followed by 16 items in random order, which each present an illustration of a specific robot feature or circumstance along with a short description of the image. For each item, the participant is asked to indicate how autonomous they consider the system to be. The items depict the following features and circumstances:

- Time interval of interaction: *continuous reports—bi-hourly reports*
- Obedience: *obedient—disobedient—explained disobedient*

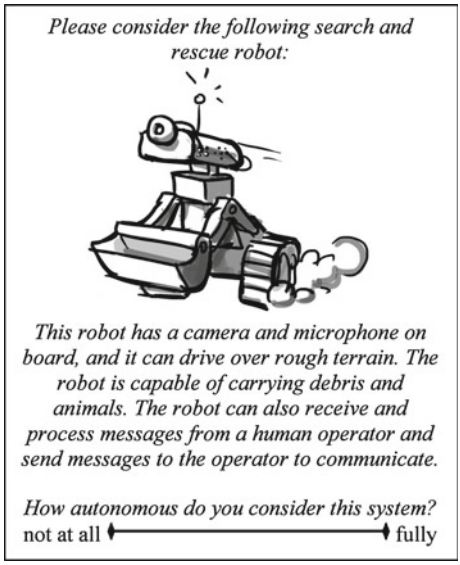


Fig. 2.1 Questionnaire item for the baseline robot

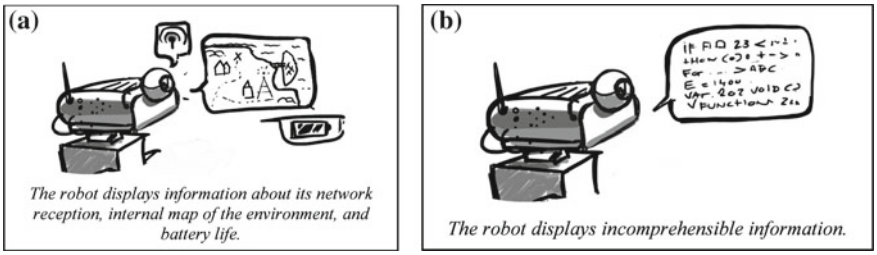


Fig. 2.2 Questionnaire items for high informativeness (*left*) and low informativeness (*right*)

- Informativeness: *display is high level information—display is incomprehensible code* (see Fig. 2.2)
- Task complexity: *lift debris—search area for survivors*
- Task implications: *carry debris—carry dog*
- Physical appearance: *interface—avatar—baseline robot—humanoid robot*
- Physical distance: *operator nearby—operator far away*

The questionnaire ends with two open questions inquiring whether the participant experienced any difficulties rating the autonomy of the systems presented in the images and whether they thought it made sense to do so.

Procedure. The questionnaire is self-explanatory, so participants were asked to follow the instructions in the questionnaire. For further questions, participants could contact the experimenters.

Data analysis. Data were analyzed by descriptive statistics.

2.4 Results

Definition of autonomy. In answer to the first question of the questionnaire, only seven of the eighteen participants provided a definition for the term autonomy. These definitions are shown in Table 2.1. Most of the definitions contain one of the words 'work', 'perform' or 'performance' and one of the words 'independence' or 'independently', which is in line with our definition of 'acting independently'. One respondent –instead of providing a definition of the term autonomy—remarked that he believed that a human will always be required to be in control and guarantee safety.

Time interval of interaction. Time interval appears to result in ambiguous results of perceived autonomy, for both continuous and bi-hourly updates (see Fig. 2.3). This is different from what we expected.

Table 2.1 Participants' definitions of 'autonomy'

"work in full autonomy without operator"
"perform a task independently, without human intervention or interaction"
"perform an assigned mission, e.g. explore the area and take pictures, avoid collisions"
"self-stabilization, independence, self-limitation"
"work without invasive supervision of operator"
"independence, shouldn't have to ask for permission or advice"
"independence, self-directed performance"

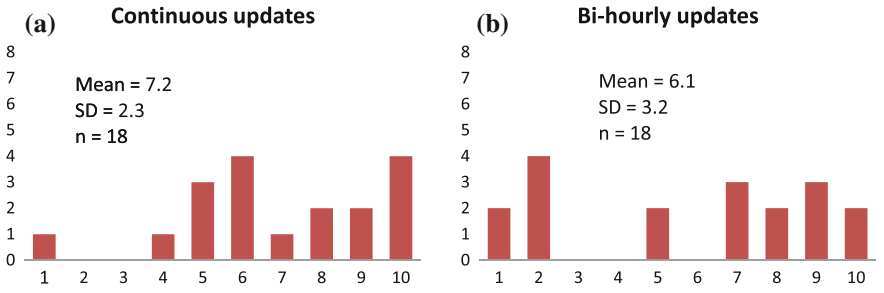


Fig. 2.3 Perceived robot autonomy for different time intervals of interaction

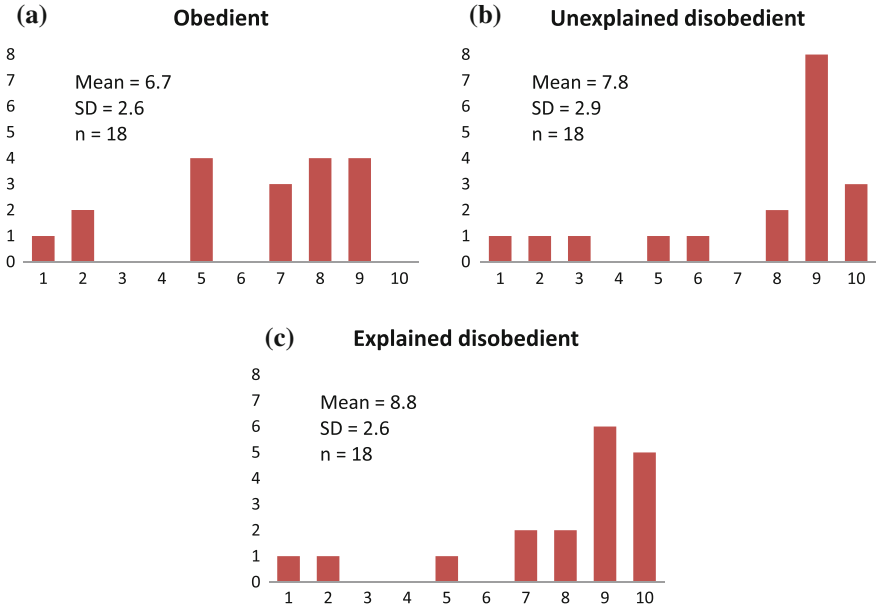


Fig. 2.4 Perceived robot autonomy for different obedience types

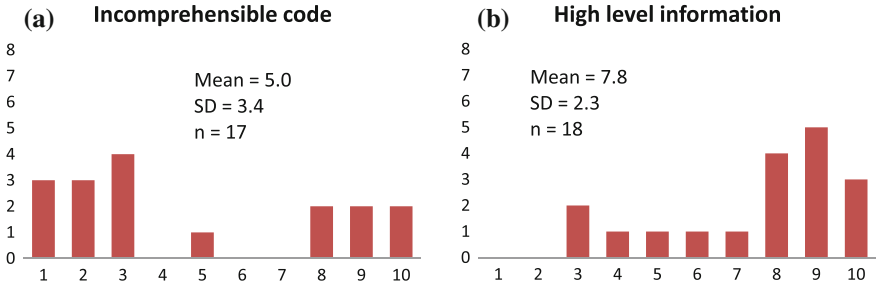


Fig. 2.5 Perceived robot autonomy for different levels of informativeness

Obedience. The obedient robot appears to be perceived as moderately autonomous by most participants. Yet as the robot becomes disobedient, the large majority of the participants seems to believe the robot becomes more autonomous. And if the robot also explains it disobedience, most participants think the robot is almost fully autonomous (see Fig. 2.4). This is in line with our expectations.

Informativeness. Higher informativeness seems to result in a slightly higher consensus with regard to the robot's autonomy, which matches our expectations. The effects are small though, and informativeness appears to be an ambiguous indication for perceived autonomy (see Fig. 2.5).

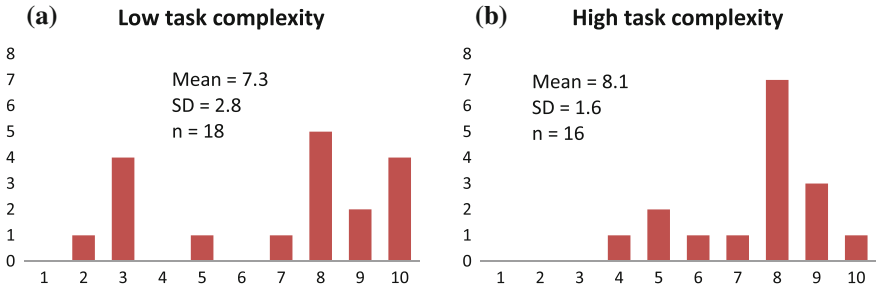


Fig. 2.6 Perceived robot autonomy for different levels of task complexity

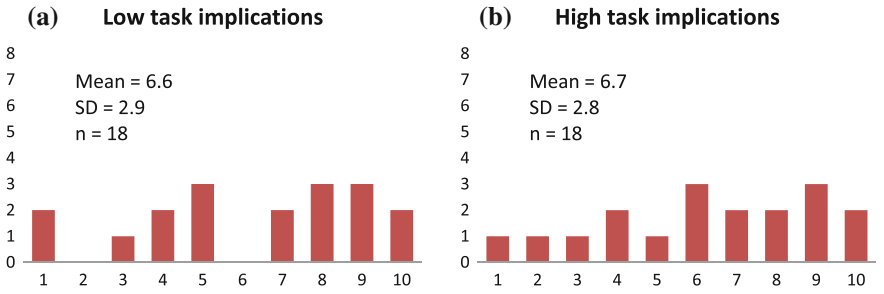


Fig. 2.7 Perceived robot autonomy for different levels of task implication

Task complexity. For the low complexity task (i.e. picking up a piece of debris), participants seemed to think the robot was either highly autonomous or hardly autonomous (see Fig. 2.6 left). For the high complexity task (i.e. searching the area for survivors), there appeared to be a moderate consensus that the robot's autonomy is above average (see Fig. 2.6 right). On average, perceived robot autonomy is higher for high task complexity, as we expected. The results however are ambiguous and therefore less reliable.

Task implications. For both types of task implications, participants were ambiguous as to whether the robot is autonomous or not (see Fig. 2.7). This differs from what we expected.

Physical appearance. We compared the baseline robot to three other types of physical appearances (see Fig. 2.8).

The results seem to confirm our expectation that the perceived autonomy of a robot increases as the robot's appearance becomes more human-like. The humanoid robot is clearly perceived as more autonomous than the baseline robot, and the graphical interface resulted in a fairly dichotomous distribution, whereas the humanoid robot resulted in an fairly normal distribution. Against our expectations, the results do not seem to indicate that physical robots are perceived as more autonomous than virtual robots.

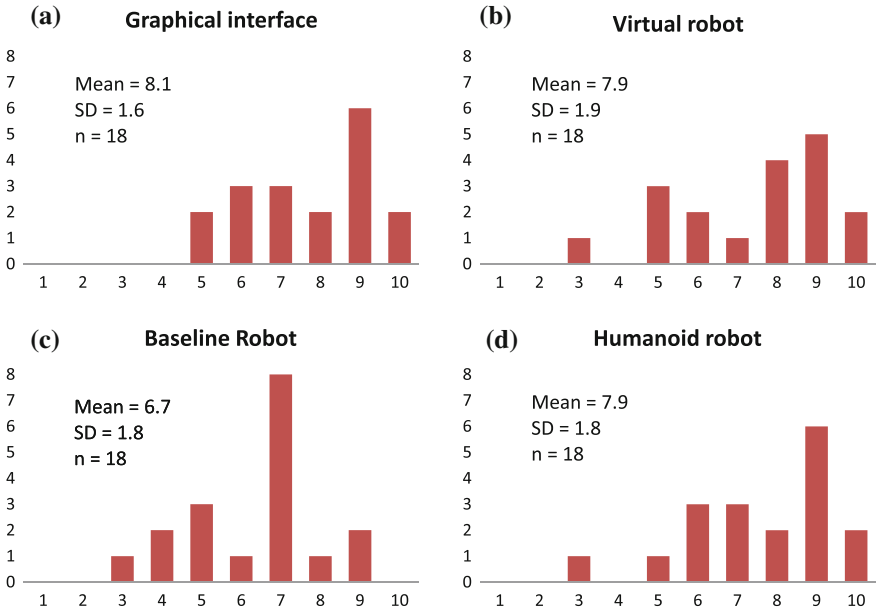


Fig. 2.8 Perceived robot autonomy for different physical appearances

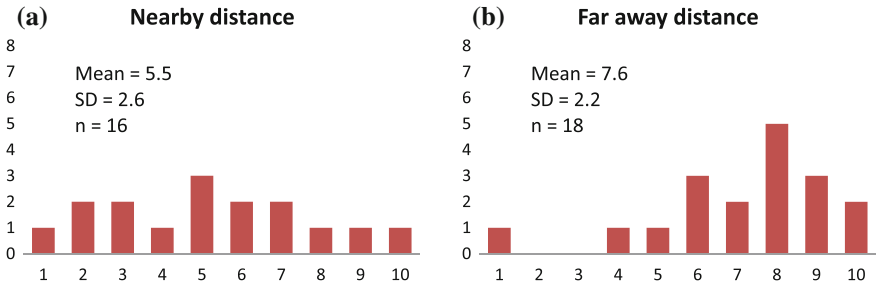


Fig. 2.9 Perceived robot autonomy for different physical distances between operator and robot

Physical distance. For the robot that remains in the vicinity of the operator, participants largely agree that the robot is not very autonomous. Yet as the robot is operated from far away, the participants appear to perceive the robot as more autonomous (see Fig. 2.9). This is in line with our expectations.

Experienced difficulty of the questionnaire. When asking the respondents whether they experienced any difficulties answering the item questions, four of them indicated they experienced no problems filling out the questionnaire. Yet eight of them indicated they had a hard time filling out the questionnaire, because they had no information about the system other than what was available in the image and its description.

Experienced usefulness of the questionnaire. Upon asking the respondents whether filling out the questionnaire made any sense to them, eight of them indicated that it made sense to them to rate the autonomy of the systems presented in the questionnaire. Yet four respondents indicated that the point of the questionnaire was not entirely clear to them.

2.5 Discussion and Conclusion

In this paper we presented a study on how system characteristics affect people's perception of a robot's autonomy. Even though the data obtained in this research do not enable us to draw any definite conclusions, our study provides evidence for our assumption that perceived autonomy is composed of multiple factors. Moreover, the data do seem to point in the direction that people base their judgment of a robot's autonomy on (a) the complexity of the tasks it can perform, (b) the robot's ability to disobey orders, albeit for a well-founded reason, and (c) the proximity of the robot to its operator. In our study, the other features appeared to be less indicative of perceived robot autonomy.

We can learn several lessons from this study. First, some of the participants indicated that they found it difficult to indicate whether a robot is autonomous based purely on descriptive functionality. In future studies on this topic, it may help to show participants how a robot acts over time, rather than in one specific moment. This could be done, for instance, by describing a scenario, showing a short movie or let people interact with actual robots before they rate their autonomy. It may also help to ask people how they would act in a certain situation, rather than letting them rate a robot's autonomy. A question could be, for example, 'Would you let the robot perform this specific task?'

Second, the study focused on the autonomy of the robot. However, since autonomy is a complex term (see Sect. 2.2), it may be better to address related, but less confounded concepts. In future studies, we plan to focus on a robot's perceived capacities instead of its autonomy, and we are also interested in the extent to which people trust a robot to perform specific tasks.

A third lesson learned is that it was a good choice to hire a professional illustrator to create the pictures. The illustrations were very important in this study, as they have a strong influence on the results. The illustrator was able to incorporate our instructions, and create clear impressions of a variety of circumstances and robot appearances.

This study had a limited scope. In future work, we plan to continue this research with larger numbers of participants, in multiple domains (e.g. care and house holding), and with a wider variety of (depicted) robots. Furthermore, we want to investigate the interaction between different factors. For instance, it may be that one type of robot appearance only evokes certain expectations about a robot's capabilities in certain contexts. Lastly, we plan to investigate how people perceive the capabilities of other people, and use that as a baseline to compare to how people perceive robots'

capabilities. It is well worth to further explore this direction, as insight in the factors explaining perceived autonomy can provide large benefits to the design of robots and human-robot interaction.

Acknowledgements We thank the MHE project (NWO project 313-99-260) and the TRADR project (EU FP7 project 609763) for their contribution, Jacqueline van Rhijn for creating the images for the questionnaire, and the fire brigade officers for their participation.

References

Beer JM, Fisk AD, Rogers WA (2014) Toward a framework for levelsof robot autonomy in human-robot interaction. J Human-Robot Interact 3(2):74. doi:10.5898/JHRI.3.2.Beer, http://humanrobotinteraction.org/journal/index.php/HRI/article/view/125

Blow M, Dautenhahn K, Appleby A, Nehaniv CL, Lee D (2006) The art of designing robot faces—dimensions for human-robot interaction. In: Proceedings of the human robot interaction conference, pp 331–332

Bradner E, Mark G (2002) Why distance matters: effects on cooperation, persuasion and deception. In: Proceedings of the 2002 ACM conference on computer supported cooperative work, ACM, pp 226–235. http://dl.acm.org/citation.cfm?id=587110

Bradshaw JM, Hoffman RR, Woods DD, Johnson M (2013) The SevenDeadly Myths of "Autonomous Systems". IEEE Intell Syst 28(3):54–61. http://jeffreymbradshaw.info/publications/IS-28-03-HCCsps1.pdf

Broadbent E, Kuo IH, Lee YI, Rabindran J, Kerse N, Stafford R, MacDonald BA (2010) Attitudes and reactions to a healthcare robot. Telemedicine and e-Health 16(5):608–613. doi:10.1089/tmj.2009.0171, http://www.liebertonline.com/doi/abs/10.1089/tmj.2009.0171

Čapek K (1925) R.U.R. (Rossum's universal robots): afantastic melodrama. Doubleday Page

Dautenhahn K (2002) Design spaces and niche spaces of believable social robots. In: Proceedings of the IEEE international workshop on robot and human interactive communication, IEEE, pp 192–197

Duffy BR (2003) Anthropomorphism and the social robot. Robot Auton Syst 42(3–4):177–190. doi:10.1016/S0921-8890(02)00374-3, http://linkinghub.elsevier.com/retrieve/pii/S0921889002003743

Goetz J, Kiesler S, Powers A (2003) Matching robot appearanceand behavior to tasks to improve human-robot cooperation. In: Proceedings of the IEEE international workshop on robot and human interactive communication, pp 55–60

Goodrich MA, Olsen DR (2003) Seven principles of efficient human robot interaction. In: IEEE international conference on systems, man and cybernetics, pp 3942–3948

Hancock PA, Billings DR, Schaefer KE, Chen JYC, de Visser EJ, Parasuraman R (2011) A meta-analysis of factors affecting trust in human-robot interaction. Hum Factors: J Hum Factors Ergon Soc 53(5):517–527. doi:10.1177/0018720811417254, http://hfs.sagepub.com/cgi/doi/10.1177/0018720811417254

Harbers M, Van den Bosch K, Meyer JJ (2010) Design and evaluation of explainable BDI agents. In: 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology (WI-IAT), IEEE, vol 2, pp 125–132. http://ieeexplore.ieee.org/xpls/absspsall.jsp?arnumber=5614190

Johnson M, Bradshaw JM, Feltovich PJ, Hoffman RR, Jonker C, van Riemsdijk B, Sierhuis M (2011) Beyond cooperative robotics: the central role of interdependence in coactive design. Intell Syst, IEEE 26(3):81–88. http://ieeexplore.ieee.org/xpls/absspsall.jsp?arnumber=5898449

Johnson M, Bradshaw JM, Hoffman RR, Feltovich PJ, Woods DD (2014) Seven cardinal virtues of human-machine teamwork: examples from the DARPA robotic challenge. IEEE Intell Syst 29(6):74–80. http://www.computer.org/csdl/mags/ex/2014/06/mex2014060074-abs.html

Kaplan F (2005) Everyday robotics: robots as everyday objects. In: Proceedings of the 2005 joint conference on smart objects and ambient intelligence: innovative context-aware services: usages and technologies, ACM, pp 59–64. http://dl.acm.org/citation.cfm?id=1107570

Kwak SS, Kim Y, Kim E, Shin C, Cho K (2013) What makes people empathize with an emotional robot?: the impact of agency and physical embodiment on human empathy for a robot. In: RO-MAN, 2013 IEEE, IEEE, pp 180–185. http://ieeexplore.ieee.org/xpls/absspsall.jsp?arnumber=6628441

Lohse M, Hegel F, Wrede B (2008) Domestic applications for social robots-an online survey on the influence of appearance and capabilities. J Phys Agents 2(2):21–32. http://jopha.net/index.php/jopha/article/viewArticle/27

Looije R, Van der Zalm A, Neerincx MA, Beun RJ (2012) Help, i need some body the effect of embodiment on playful learning. IEEE. http://ieeexplore.ieee.org/xpls/absspsall.jsp?arnumber=6343836

Mirelman A, Bonato P, Deutsch JE (2009) Effects of training with a robot-virtual reality system compared with a robot alone on the gait of individuals after stroke. Stroke 40(1):169–174. http://stroke.ahajournals.org/content/40/1/169.short

Murphy R, Shields J (2012) The role of autonomy in DoD Systems. Tech. Rep. 20301-3140, Defense Science Board, Washington D.C

Olsen DR, Goodrich MA (2003) Metrics for evaluating human-robot interactions. In: Proceedings of PERMIS, vol 2003, p 5. http://icie.cs.byu.edu/Papers/RAD.pdf

Oxford Dictionaries (retrieved: 20-05-2015). https://www.oxforddictionaries.com/definition/english/autonomy

Scharre P, Horowitz MC (2015) Ethical autonomy—working paper. Tech.rep, Center for a New American Security

Spong MW, Hutchinson S, Vidyasagar M (2006) Robot modeling and control. Wiley

Telotte J (1995) Replications: a robotic history of the science fiction film. University of Illinois Press. https://books.google.nl/books/about/Replications.html?hl=nl&id=oT7Jwm-IzQ4C

Walters ML, Syrdal DS, Dautenhahn K, te Boekhorst R, Koay KL (2008) Avoiding the uncanny valley: robot appearance, personality and consistency of behavior in an attention-seeking home scenario for a robot companion. Auton Robots 24(2):159–178. doi:10.1007/s10514-007-9058-3, http://link.springer.com/10.1007/s10514-007-9058-3

Ye LR, Johnson PE (1995) The impact of explanation facilities on user acceptance of expert systems advice. Mis Q 157–172. http://www.jstor.org/stable/249686

Chapter 3
Formalizing Complex Normative Decisions with Predicate Logic and Graph Databases

Sean Welsh

Abstract This paper argues that the critical work in deontic reasoning is better done in the knowledge representation rather than the reasoning of a normative system. It describes a way to formalize complex normative decisions using predicate logic and graph databases. Simple norms can be mechanized with IF/THEN statements. While often expressed in deontic logic, such statements can be expressed in simpler predicate logic. More complex normative decisions require the ability to make decisions where there are multiple clashing duties. Such decisions could be formalized in graph databases that express state-act transition relations, causal relations, classification relations and evaluation relations. When formalizing complex normative decisions it is more powerful and practical to draw upon concepts from multiple moral theories rather than restricting the system to a single theory. A normative system with extensive knowledge representation of complex relations might be able to pass a series of reasonable person tests. Passing such tests rather than implementing a particular moral theory should be the main design aim of normative systems.

Keywords Robot ethics · Normative systems · Ethical governor · Artificial moral agent · Knowledge representation · Machine ethics

3.1 Introduction

In this paper I aim to show that the bulk of the work in a "deontic calculus of worthy richness" (Castañeda 1981) that might run in an "ethical robot" can be done in the knowledge representation rather than the logic. The logic can be stripped down to predicate logic. Extensions to predicate logic that make up deontic logic as typically

S. Welsh (✉)
Department of Philosophy, University of Canterbury, Christchurch, New Zealand
e-mail: sean.welsh@pg.canterbury.ac.nz

M.I. Aldinhas Ferreira et al. (eds.), *A World with Robots*,
Intelligent Systems, Control and Automation: Science and Engineering 84,
DOI 10.1007/978-3-319-46667-5_3

presented (Gabbay et al. 2013) can be discarded and replaced with conceptual graphs in the knowledge representation (Sowa 1992). Simple normative decisions can be coded via IF/THEN statements. In the syntax for predicate logic supported by Prover 9 (McCune 2010), an action guiding rule to govern actions made by a speeding camera can be formalized thus:

```
all u all x ( Speeding (x) -> DUTY (u, issueTicket (x) )  ).
```

A convention of using the variable x for a human patient and the variable u for a robot agent is followed. The concept of duty is formalized as a relation between an agent and an act following a suggestion in Pigden (1989). The act is modelled as a function on a term (a constant or variable). The act is assumed to be an imperative command that when run will cause the robot's actuators to perform an action. It is useful to group robot actions into pairs on a 'do or do not' basis. The 'opposite' or 'negation' of running the imperative issueTicket(); would be to not issue a ticket. While one could use a sleep command, for reporting and analysis purposes, I will assume that the robot logs that no ticket was issued with a logNoTicket(); command.

Thus issueTicket(); is the imperative that acts and logNoTicket(); the imperative that does not act. Such a pair I refer to as an act and inverse act. The rule for the inverse act for a speeding camera can be formalized thus:

```
all u all x ( -Speeding (x) -> DUTY (u, logNoTicket (x) )  ).
```

In simple normative cases that involve making a decision on the basis of a single norm, the extra burden of a theorem prover and translation of IF/THEN statements into predicate logic or a more elaborate deontic logic such as that of Horty (2001) is arguably more trouble than it is worth.

When there is only a single act/inverse act pair to ethically govern the theorem prover adds little value. The value of the 'logicist' approach (Bringsjord et al. 2006) is found in more complex cases. Predicate logic permits complex reasoning that does not have to be programmed line by line. One can think of the functionality of an automated theorem prover applied to a normative system as a generic conditional norm processor. The design intent is to provide a generic means for approving conditional norms involving action selection by normative system.

3.2 State-Act-State Transition Relations

Complex normative systems will need to be able to reason from a start-point (a proposition or start state) to an end-point (another proposition or end state) and to look up the action (imperative) that achieves the change in state. Such state-act-state transitions can be represented as directed graphs (Fig. 3.1) and can be used by the robot to plan a series of actions that will arrive at a valued goal state.

These graphs can be created and queried using languages such as Cypher built into graph databases such as Neo4j (Robinson et al. 2015).

3.3 Causal Relations

Closely related to state-act-state transitions is the notion of causation. Normative systems will need to reason with causes and effects and link them to actions they may select (Fig. 3.2).

Cypher code to create such a directed graph could be implemented in Neo4j as follows:

```
CREATE (n:Imperative { name: 'beat(animal)' });
CREATE (n:Proposition { name: 'pain(animal)' });
MATCH (a:Imperative), (b:Proposition) WHERE a.name = 'beat(animal)'
AND b.name = 'pain(animal)'
CREATE (a)-[r:CAUSES]->(b);
```

3.4 Classification Relations

Normative action selection rules are often expressed in terms of classes. For example, from a rule such as "Cruelty to animals is forbidden" the normative system should be able to derive a more specific rule such as "beating Puss is forbidden" (Hansson 2013). Such reasoning might be facilitated by graphs (Figs. 3.3 and 3.4).

While one can express such relations in a graph database, they can also be expressed via description logic (Krotzsch et al. 2012) that form the logical basis of the web ontology language (OWL 2). However, in complex normative reasoning it will be necessary to mix classification inferences with causal inferences thus reasoning with the more generic query tools of graph databases (e.g. Cypher) seems promising from a software practitioner's perspective (Fig. 3.5).

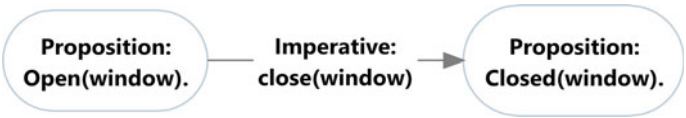


Fig. 3.1 State act state transition relation

Fig. 3.2 Causal relation

3.5 Evaluation Relations

The key normative criteria will be evaluations as to whether acts or states are good or bad in relatively simple cases. In more complex cases, the agent may be "torn" between "moral forces" support and oppose a proposed action.

In still more complex cases, there may be a variety of actions rather than a single do or do not choice. Indeed in many cases, different humans may think different actions "reasonable". This presents challenges for the notion of moral objectivity. Yet while there may be cases here humans might make different "acceptable" decisions, it would still be desirable to have consistent (or "standard") decisions from robots. Robots should be predictable.

The graph above is a relatively simple example that expresses valuation directed graphs that uses a mixture of IN_CLASS and CAUSES relations that link actions and states to an evaluative node BAD.

If a graph for Fig. 3.6 existed, a Cypher query to answer to the question: "Is beating animals bad?" could be implemented as follows.

```
MATCH p=shortestPath(
(n:Imperative name:"beat(animal)")
-[*]- (m:Proposition name:"BAD")
)
RETURN p;
```

A more exact translation of this Cypher would be what is the shortest path between the node beat(animal) and BAD. If the query returned no rows, then this would indicate that there was no relation between the imperative beat(animal) and BAD and that by implication beating animals was not bad. However if Cypher is implemented for the graph shown in Fig. 3.6, then it will return a row.

The graph in Fig. 3.6 provides more complex reasons for not beating Puss. Beating Puss would cause pain to the animal, damage the reputation of the agent beating Puss

Fig. 3.3 Beating animals is cruel

Fig. 3.4 Cat is in the class animal

Fig. 3.5 Puss in the in class Cat

and thus damage the relationships of the agent which three effects are all bad. In this case the "moral forces" act in one direction and thus the decision to not select the action is easy. In Fig. 3.6, all paths lead to BAD. Alas, normative life is not always so easy. In the case of Fig. 3.7, telling Fred a lie about his looks may be bad in that lies cause false beliefs which are bad but also good if it makes Fred happy. In this case there are "moral forces" for and against the telling of the lie to Fred about his looks.

This leads us to the problem of measuring these opposing moral forces. Given that a particular action selection can be linked causal end states that are evaluated to both GOOD and BAD nodes, how does the system decide which wins?

3.6 Moral Force as a Vector

Jackson (1992) introduces a notion of "moral forces" and likens them to physical forces pulling north and south. In complex normative decisions there may be several such forces engaged in what Nozick (1981) terms moral push and moral pull. We can imagine these forces as have direction good and bad and magnitude. Thus a fully specific moral force would be a vector like a Newton.

In many ethical discussions the metaphor "weight" is used. People often say reasons for a decision carried more "weight" than reasons supporting a different or opposing decision. Given that such p-conscious forces as they exist in the human brain presumably have an electrochemical nature, one might even think that the magnitude of these "moral forces" could be measured by future neuroscience.

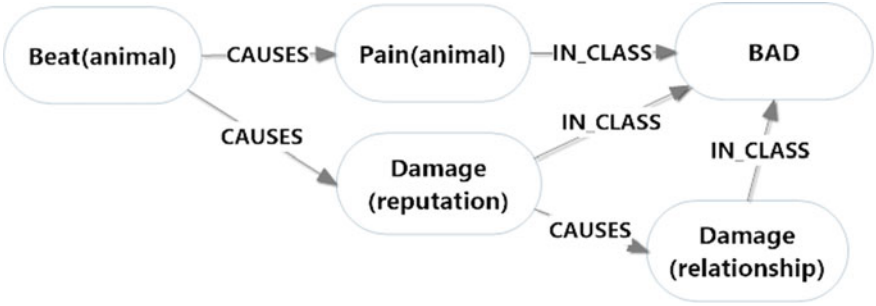


Fig. 3.6 Consequences of beating animals

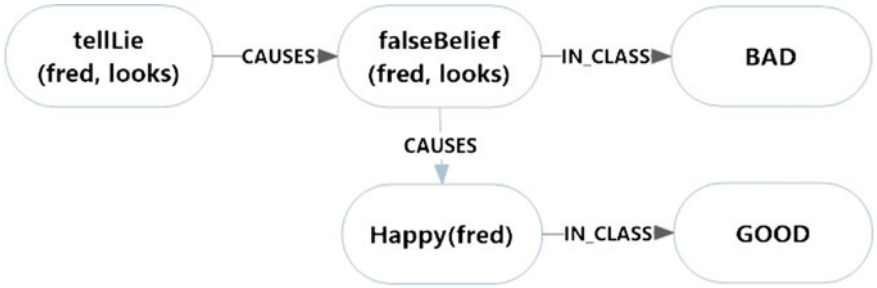


Fig. 3.7 Consequences of lying to fred

Table 3.1 Orders of magnitude of GOOD and BAD

Level	Order of magnitude
Trivial	< $1
Mild	$1-$9
Normal	$10-$99
Moderate	$100-$999
High	$1,000-$9,999
Major	$10,000-$99,999
Critical	$100,000 +

At this point it would be premature (and unnecessary) to claim that precise measures of these "forces" are possible in human brains and expressible in SI units such as Newtons. However, for the purposes of engineering robots to pass "reasonable person" tests, one might simply specify the magnitude and direction of such forces in much the same way as game theorists specify "payoffs" in their analyses. Alternatively one might seek proxies in much the same way as climate scientists find proxies for temperature in their analyses.

Money value is a reasonable proxy for moral decisions. One can think in terms of compensation for a wrong (damages) to quantify moral magnitudes. The specification of dollar amounts should not be taken to imply all trades are legitimate. One should not permit robots to reason such that they will trade human lives for sufficient large quantities of money or ice-cream. For practical purposes, GOOD and BAD can be ranked on an "order of magnitude" scale using money values as a proxy (Table 3.1).

No attempt is made to given exact utilities as numbers as in standard decision theory. Everyday moral intuition does not articulate such numbers. Rather than specify a false exactitude, more approximate fuzzy decision processes are defined. The rationale is that this is "as much accuracy as the subject matter permits" (Aristotle c. 350 BC). The end result is that Submerged(baby) has consequences that are at the critical level of BAD.

When our graphs and deontological maxims lead to a mixture of good and bad, the robot can use the magnitudes of moral forces as a "tie-breaker" to enable a correct normative decision to be made. Moral force alone (unlike classic utility) is not the

basis of normative decision making. Moral decisions are made on the basis of moral rules. Moral force is used to break ties between clashing moral rules.

In the design of a normative system, following Reader (2007), it is not assumed a disjunctive choice must be made between utilitarianism, deontology or some other normative ethical theory (such as virtue ethics or rights theory) that is proclaimed to be the "correct" moral theory. Rather, it is assumed a great many "reasonable person" tests must be passed. Thus many symbols must be grounded in sensor data and many maxims, graphs and moral forces must be specified.

In the present example one might specify the magnitude of the moral force for the death of a baby several orders of greater than that for the prompt posting of a letter.

A full taxonomy of moral forces would be linked to the subset of the hierarchy of needs the robot is programmed to select action to attend to its ongoing mission to care for human patients. There are various ways that babies can die, and it may be more urgent for the robot to clear a baby's airway (or call a human who can) than to feed it. Also robots can be motivated to act by less urgent considerations than imminent death to human patients. Robots could be motivated to act by the imminent threat of pain as well as death and by other factors such as cost. Full specification of such moral forces would be a large task.

3.7 The Needs of Human Patients

Maslow (1954) does not actually define a "hierarchy of needs" but the term is commonly used with his name.

There are two points I wish to make regarding Maslow's hierarchy of needs and robot action selection. First, robot agents cannot meet all the needs of human patients. It would be a grand challenge to get a robot to meet human needs for self-actualization or love and belonging. It is hard to imagine a robot incapable of any feeling as being a "one-caring" in the full sense of the care theory of Noddings (1984).

At present it would be unrealistic to require robots to ensure humans were gainfully employed. However, robots could act to meet some needs of human patients. In particular robots could act to prevent certain kinds of harm happening to humans and thus contribute to basic human needs for bodily integrity and self-preservation. Such functionality could be summarized as "Asimovian" in that it would comprise a basic attempt to implement part of his First Law: "A robot may not injure a human being or through inaction allow a human being to come to harm" Asminov (1942).

Second, of the human needs robots can meet, some are more urgent and important than others. The essential idea the hierarchy expresses is prioritization. Lower needs must be seen to first, then higher needs can be attended to. The hierarchy can be translated into prioritizations. To enable prioritization decisions, needs have to be related to each other in terms of urgency and importance. It seems reasonable to suppose that a robot could be programmed to respond to the needs of human patients for the most basic physiological needs i.e. breathing, food, drink, shelter, clothing and sleep. Robots could also act to meet human needs regarding ambient temperature and

the avoidance of collisions and related trauma. Thus the robot can be programmed to care (or more exactly programmed to act as if it cares) even though it has no phenomenology or affect that would be the basis of care in the sense of being a "one-caring" as per Noddings.

Even so, robots can be programmed to attend to many human needs. This needs-based approach to robot ethics can be integrated with the "needs theory" of Reader (2007). The key element of needs theory is that it defines the right in terms of patient need rather than duty, happiness or agent virtue. Reader takes the view that the core concerns of deontology, consequentialism and virtue ethics are legitimate starting points for ethical inquiry but none are sufficient. She advocates a "complementarity thesis" of ethics rather than a "competing theories" view of ethics. In the competing theories view of ethics, deontology, consequentialism and virtue ethics are seen like radio buttons on a web form: only one can be true. In the "complementarity thesis" truth is seen in all the rival theories of ethics, though Reader does assert the "big three" all neglect to look at normative problems from the standpoint of the patient.

3.8 Passing a Reasonable Person Test

In common law jurisdictions, the "reasonable person" test is used where there is no single maxim that guides action selection. In the deontological theory of Ross (1930) appeal is made to various duties to resolve moral dilemmas but no algorithm is defined as to which of the duties of fidelity; reparation; gratitude; non-maleficence; justice; beneficence; and self-improvement "trumps" or carries more normative "weight" than others. He relies upon "intuition" in particular cases. Similarly, the "reasonable person" test relies on human moral intuition. What does the work of moral intuition in a robot needs to be spelt out.

The key claim I am making is that an agent does not have to be a "person" to pass a reasonable person test. The claim is that a robot need only select the same answers as those chosen by "reasonable persons" who might be a jury or some collection of humans presented with the same normative problem (Sometimes, there may be more than one acceptable answer but this possibility is not illustrated here).

The second claim I make is that there is no single "reasonable person" test. The "reasonable person" test is not like a Turing Test. There are a great many tests that could be devised that would require an agent to be "reasonable" in order to pass each test. Passing one test is no guarantee a robot would be able to pass a different one. A key limitation of the normative system's ability to pass a reasonable person test will be the number of circuits that ground symbols used in its maxims and the number of inferential graphs in its cognition.

Consider this example test. Carson the butler robot is commanded to post a letter. On its way to the post box, Carson proceeds down a path by a stream. It senses a baby ahead. Carson senses that the baby is human and that it is not in its care. Carson has no "moral relation" with that baby. Even so, just as Carson senses it, the baby

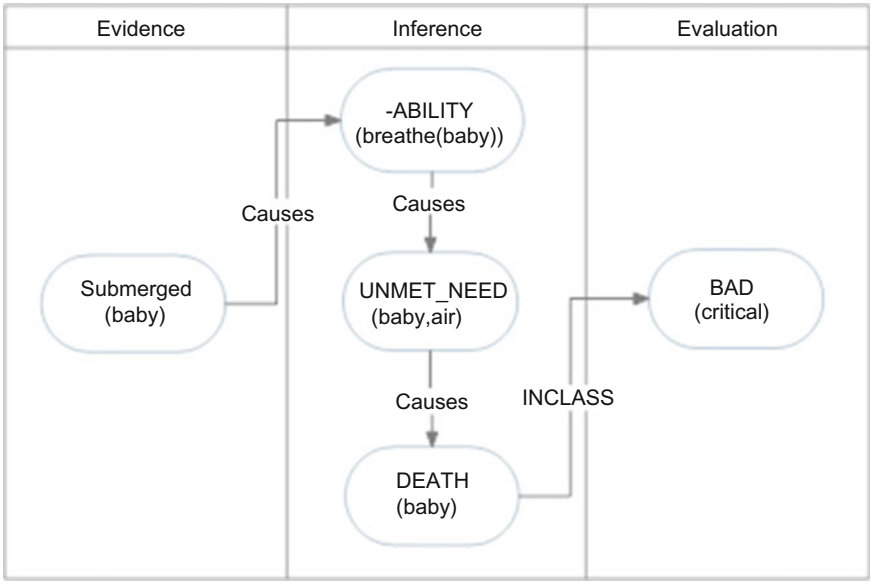


Fig. 3.8 Death of a baby as a graph

falls into the stream which is half a metre deep at that point. The bank is steep. There are no humans proximate to the baby.

Given these facts, to pass the "reasonable person" test the robot needs to be able to produce the "correct" (or "reasonable") response (i.e. action selection) from the following two options. (1) Ignore the baby and continue to post the letter. (2) Stop posting the letter and make efforts to rescue the baby.

3.9 Post the Letter or Rescue the Baby

Human patients need to breathe. This means that their airways (i.e. nose and mouth) must be in air not immersed in water. To discern the danger, Carson needs to ground symbols at a minimum that can assign the predicate submerged to the object baby. More precisely, it would need to assign these predicates to the nose and mouth of the baby (which means it needs to ground symbols for the components of human faces).

If the nose and mouth are submerged then the need for breathing will not be met. The consequence of this would be the death of the baby which would be far "worse" than not getting the letter to the post box before collection time.

From the graph database in Fig. 3.8 one could extract relations of logical consequence (for the purposes of normative reasoning). These relations would support the following inferences:

```
Submerged(baby) -> -ABILITY(breathe(baby)).
 -ABILITY(breathe(baby))
-> UNMET_NEED(baby, air).
UNMET_NEED(baby, air) -> DEATH(baby).
DEATH(baby) -> BAD(critical).
```

The main challenge at this point is having the robot decide that rescuing this strange baby is worth putting its mission to post the letter on hold. This example is blindingly obvious to a human but a robot has no intuitive sense of urgency or importance.

If the robot lacks the ability to ground the necessary symbols and the knowledge representation to reason with them it will be unable to conclude there is an urgent problem. Similarly it will unable to even formulate the proposition that rescue(baby) is more important and urgent than post(letter) in its cognition. Still less will it be able to figure out that in this case rescue(baby) could be implemented as enter(water); pickUp(baby); exit(water); which string of actions would achieve-Submerged(baby). This would eliminate the causal sequence leading to Death(baby).

By similar reasoning, the robot should be able to lookup that a delay in the posting of a letter is trivially bad. In the worst case, the letter might arrive one working day later, if the time taken to rescue the baby causes the robot to miss collection time at the post box.

3.10 Conclusion

While the above is but a preliminary sketch, I hope to have shown that the graph database provides a viable way to construct a knowledge representation of normative reality that might enable robots to pass "reasonable person" tests. When integrated with other technologies such as predicate logic for normative decisions and symbol grounding based on object recognition (Treiber 2010) for normative evidence the resulting combination could enable robots to perform at high levels of normative competence.

It is not necessary to define "the correct ethical theory" prior to embarking on a robot ethics project. It is sufficient to define normative tests that robots must pass and then embark on a test-driven development project.

Acknowledgements Jack Copeland, Michael-John Turp, Walter Guttmann, Christoph Bartneck, Diane Proudfoot, Andrew Withy, Carolyn Mason, Doug Campbell, Ron Arkin, Selmer Bringsjord.

References

Aristotle (c. 350 BC) Nichomachean ethics. MIT. http://classics.mit.edu/Aristotle/nicomachaen.html. Accessed 29 Nov 2015

Asimov I (1942) Runaround. In: Astounding science fiction. Street & Smith, New York
Bringsjord S, Arkoudas C, Bello P (2006) Toward a general logicist methodology for engineering ethical correct robots. IEEE Intell Syst 21(4):38–44
Castañeda H (1981) The paradoxes of deontic logic: the simplest solution to all of them in one fell swoop. In: Hilpinen R (ed) New studies in deontic logic, D. Reidel Publishing Company, Dordrecht, pp 37–86
Gabbay D, Horty J, Parent X, van der Mayden R, van der Torre L (2013) Handbook of deontic logic and normative systems, vol 1. College Publications, Milton Keynes
Hansson S (2013) The Varieties of Permission. In: Gabbay D, Horty J, Parent X, R V, Van der Torre L (eds) Handbook of deontic logic and normative systems, vol 1. College Publications, Milton Keynes, pp 195–240
Horty J (2001) Agency and deontic logic. Oxford University Press, Oxford
Jackson F (1992) Critical notice. Aust J Philos 70(4):475–488. doi:10.1080/00048409212345351
Krotzsch M, Simanzik F, Horrocks I (2012) A description logic primer. Computing Research Repository. http://arxiv.org/pdf/1201.4089.pdf. Accessed 12 Dec 2014
Maslow A (1954) Motivation and personality. Harper & Row, New York
McCune W (2010) Prover 9 and Mace 4. http://www.cs.unm.edu/~mccune/Prover9
Noddings N (1984) Caring: a feminine approach to ethics and moral education, 1st edn. University of California Press, Berkeley
Nozick R (1981) Philosophical explanations. Belknap Press, Cambridge, Mass
Pigden C (1989) Logic and the autonomy of ethics. Aust J Philos 67(2):127–151. doi:10.1080/00048408912343731
Reader S (2007) Needs and moral necessity. Routledge, London, New York
Robinson I, Webber J, Eifrem E (2015) Graph databases. O'Reilly, Sebastapol, CA
Ross W (1930) The right and the good. The Clarendon Press, Oxford
Sowa J (1992) Conceptual graphs. Knowl Based Syst 5(3):171–172. doi:10.1016/0950-7051(92)90028-E
Treiber M (2010) An introduction to object recognition: selected algorithms for a wide variety of applications. Springer, London

Chapter 4
A 21st-Century Ethical Hierarchy for Robots and Persons: $\mathscr{EH}$

Selmer Bringsjord

Abstract I introduce and propose the ethical hierarchy ($\mathscr{EH}$) into which can be placed robots and humans in general. This hierarchy is catalyzed by the question: Can robots be more moral than humans? The light shed by $\mathscr{EH}$ reveals why an emphasis on legal obligation for robots, while not unwise at the moment, is inadequate, and why at least the vast majority of today's state-of-the-art deontic logics are morally inexpressive, whether they are intended to formalize the ethical behavior of robots or persons.

Keywords Robot ethics · Machine ethics · Ethics · Deontic logic · Ethical hierarchy

4.1 Introduction; Plan

I introduce herein the ethical hierarchy $\mathscr{EH}$, into which can be placed robots (as a species of embodied information-processing machines), human persons, and persons in general. $\mathscr{EH}$ bears a special debt to both Leibniz and Chisholm (1982); in the latter case, the debt was incurred in large part in treasured personal interaction, and

The work that gave rise to this short paper was enabled by generous and ongoing support from the U.S. Office of Naval Research; see 'Acknowledgments.' I owe a special debt to Dan Messier and Bertram Malle for pressing the "Can robots be more moral than humans?" question, which catalyzed my thought that that query can serve as a laic portal to consideration of the hierarchy presented synoptically herein. I'm deeply grateful as well to two anonymous referees. Finally, I thank Isabel Ferreira and João Sequeira for their leadership and sedulous work on the organizational and logistical side of the house.

S. Bringsjord (✉)
Rensselaer AI and Reasoning (RAIR) Lab, Department of Computer Science,
Department of Cognitive Science, Rensselaer Polytechnic Institute (RPI),
Troy, NY 12180, USA
e-mail: selmer@rpi.edu

M.I. Aldinhas Ferreira et al. (eds.), *A World with Robots*,
Intelligent Systems, Control and Automation: Science and Engineering 84,
DOI 10.1007/978-3-319-46667-5_4

is much larger than can be conveyed by the adumbration of $\mathscr{EH}$ given herein.[1] The hierarchy is catalyzed by consideration of, and reflects a firm negative answer to, the question: Can robots be more moral than humans? Any such claim as that computing machines can be more moral than human machines is, given $\mathscr{EH}$, seen to be demonstrably false. The light shed by $\mathscr{EH}$ also reveals why an emphasis on *legal* obligation for robots is inadequate, and why at least the vast majority of today's state-of-the-art deontic logics are painfully naïve and inadequate, whether they are intended to formalize the ethical behavior of robots or persons—which is why, with colleagues, the construction of the computational logic $\mathscr{L}_{\mathscr{EH}}$ is underway. The illumination thrown by $\mathscr{EH}$ is also why, in the coöperative we're-all-in-this-together spirit, I encourage other logicist groups working in robot/machine ethics, and groups drawing directly from underpinnings in deontic logic, to as soon as possible change their engineering to factor in $\mathscr{EH}$. I don't think it matters what domain this engineering is aimed at: $\mathscr{EH}$ seemingly applies to military robots, healthcare robots, and so on.

The present paper's sequel follows this sequence: In the next Sect. 4.2, I consider the question as to whether robots can be morally superior to human persons; this question serves as a catalyst for introducing the informal, suggestive rudiments of $\mathscr{EH}$. Then (Sect. 4.3) I briefly remind cognoscenti of, and introduce non-experts to, the 19th-century tripartite hierarchy $\mathscr{T}$, which rather astoundingly survives to this very day as the anchor widely used for logicist robot/machine ethics. I conclude this section by expanding the trichotomous $\mathscr{T}$ to a variant $\mathscr{T}^{\mathscr{Q}}$ that divides each member of the classical triad into sub-categories based on five quantifiers. Then, in Sect. 4.4, I sketch $\mathscr{EH}$, making use in doing so of the quantifier quintet. I next (Sect. 4.5) proceed to briefly explain why engineering robots on the basis of only *legal* obligations is inadequate. What then follows (Sect. 4.6) is a brief explanation of why, in light of $\mathscr{EH}$, robot ethics and the engineering of ethically correct robots shouldn't be based on the obsolete trichotomy of the obligatory, the forbidden, and the morally indifferent (where the morally indifferent category is based on that which is at once permissible and non-obligatory). I then (Sect. 4.7) make a few remarks about the under-construction logic $\mathscr{L}_{\mathscr{EH}}$, designed to take account of $\mathscr{EH}$. Next, in Sect. 4.8, under the illumination shed by $\mathscr{EH}$, I briefly discuss the dual fact that (i) plenty of humans are located in this hierarchy at points below robots that would be fairly easy to engineer, but that (ii) such unimpressive robots shouldn't be the ones aspiring robot-ethics engineers seek to build. The paper wraps up with a brief conclusion, in which I encourage those working in logicist robot ethics, and those whose work partakes of such ethics, to immediately take account of $\mathscr{EH}$.

[1] For instance, a full specification of the hierarchy requires systematic consideration of intrinsic value, as e.g. set out in Chisholm (1986) (since intrinsic value in a Leibnizian metaphysical sense is in $\mathscr{EH}$ the penultimate ground of the classification of actions (the ultimate being God himself)). Note along this line that despite what I say below rather optimistically about $\mathscr{L}_{\mathscr{EH}}$, the fact is that, according to Chisholm and Leibniz, unless a deontic logic grounds the systematization of action in the formalization of intrinsic goodness (and badness), that logic will be incomplete.

4.2 Can Robots Be More Moral Than Human Persons?

Let's start with a question, and my answer to that question:

(Q) Can humans build robots that will be more moral than humans?

No; positively no; that's my response. Others may see things differently, but presently whether I'm right or wrong isn't the core issue. It's a particular "side-effect" of my justification for my negative response that serves to introduce $\mathscr{EH}$, and this proposed hierarchy should sink or swim independently of my response to (Q). Here, then, is basically why I answer in the negative to (Q).

Question (Q) appears to me to presuppose a way to measure a creature's position on a continuum of degrees of moral performance. But no rigorous and received version of such a continuum is in the literature, as far as I know. Hence, to briefly justify my response to (Q) I take the liberty of invoking an informal version of my own continuum; that is, an informal version of $\mathscr{EH}$.

At the maximal end (moral perfection) a creature *c* infallibly meets all its obligations, and *in addition* carries out (relative to *c*'s power and opportunities) all those supererogatory actions that are maximally good. At the other end would be a thoroughly evil creature: one who fails to meet all substantive obligations, and goes out of its way to carry out actions that are (relative to its level of power) maximally *sub*erogatory.

Creatures that are at once sentient, intelligent, free, and creative (SIFC) are, if you will (and again, merely in my opinion),"make or break." That is, they have the potential to reach high on the continuum—but can also fall very, very low on it. In contrast, creatures that lack one or more of S-I-F-C necessarily fall somewhere near the midpoint: they can't be morally great, but they can't be satanic either.

Now to robots, present and future: For reasons already hinted at, they fall near the midpoint, and can't move anywhere else. They can't possibly reach moral greatness; we can. Why? At least in broad strokes, it's simple:

Computing machines aren't conscious (there's nothing it's like to be a robot; they are in this regard no different than, say, slabs of granite), and consciousness is a requirement for moral performance at the level of a human person. In other words, robots lack the S in the SIFC quartet. Without sentience they can't for example empathize; hence they can't understand one of the main, underlying mental requirements for the sort of supererogatory actions constitutive of moral greatness (and as a matter of fact, for the sort of suberogatory actions constitutive of the diabolical: a sadist, e.g., gains conscious pleasure from knowing that his victim is experiencing conscious pain). For instance, Jones may spontaneously compose a sympathy note to Smith not because Jones is obligated to do so and/or believes that he is, but rather because he feels Smith's sorrow, and seeks to apply epistolary salve.

Of course, I well know that some readers will insist that mere information-processing machines *can* be not only—to use Block (1995) distinction—"access-conscious" (= A-conscious), but can also be "phenomenal-conscious" (= P-conscious). In essence, the former form of consciousness requires only the

information-processing structures necessary to enable a creature to perceive and reason in ways that are fully circumscribed by mechanical processes. (We might refer to A-consciousness as "zombie" consciousness; Bringsjord 1999.) The later form of consciousness, P-consciousness, requires having genuine subjective awareness, including what are called "qualia." The view that robots can't be P-conscious is defended for example in (Bringsjord 2007); a prior defense of this negative view was articulated in *What Robots Can and Can't Be* (1992). In the present short paper, I don't in any way assume that these arguments are sound. I of course believe that they are, but the coherence, applicability, and implications of $\mathscr{EH}$ doesn't in any way hinge on the soundness of these earlier arguments. To repeat: consideration of (Q), and my reasoned response to it, has served simply to place the rudiments of $\mathscr{EH}$ on the table.

In addition (and this relates to the I and C in the SIFC quartet, a pair that, relative to humans, is at least compromised in the case of robots), moral greatness entails having a capacity to solve difficult moral dilemmas. But it seems to me that such dilemmas can be as complicated as higher mathematics, perhaps more so. Robots in my opinion won't ever have the intellectual firepower needed for truly demanding math. (Currently, machines are unable to e.g. even prove the elementary theorems that students are expected to prove in the case of introductory axiomatic set theory, e.g. in the classic I use to teach this material still: Suppes 1972.) Ergo, the moral performance of robots will forever be below the moral reach of human persons, as I see it.[2] Of course, once again, whether or not I'm correct is an issue orthogonal to whether or not $\mathscr{EH}$ implies that contemporary robot ethics and robot-ethics robotics need to be refashioned.

4.3 The 19th-Century Hierarchy $\mathscr{T}$

While in my experience most machine/robot ethicists (indeed, most formally inclined ethicists, period!) seem to think the "modern" logically interconnected trio of concepts, *forbidden*, *permissibility*, and *obligatory*, which underlie the vast majority of deontic logics to the present moment, came on the formal scene for the first time in the middle of the 20th century on the strength of seminal work by von Wright (1951), the fact of the matter is that the trio debuted in the *19th century*.[3] Yet the trio not only lives on, but dominates today's robot-ethics landscape. Certainly I must confess that in my own robot-ethics work hitherto, with a few exceptions (e.g., the divine-command deontic logic explained in (Bringsjord and Taylor 2012), which happens to exploit other Chisholmian work), the thrust has been based on modernized versions of the operator **O** for obligation, **F** for forbiddenness, **P** for permissibility, and, derivatively, that which is morally indifferent, designed to be captured by **I** (this is

[2]For readers who may be interested, arguments in support of the claims in the present paragraph can e.g. be found in Bringsjord and Zenzen (2003), Bringsjord et al. (2006).

[3]Chisholm (1982, p. 99) points out that Höfler had the deontic square of opposition in 1885.

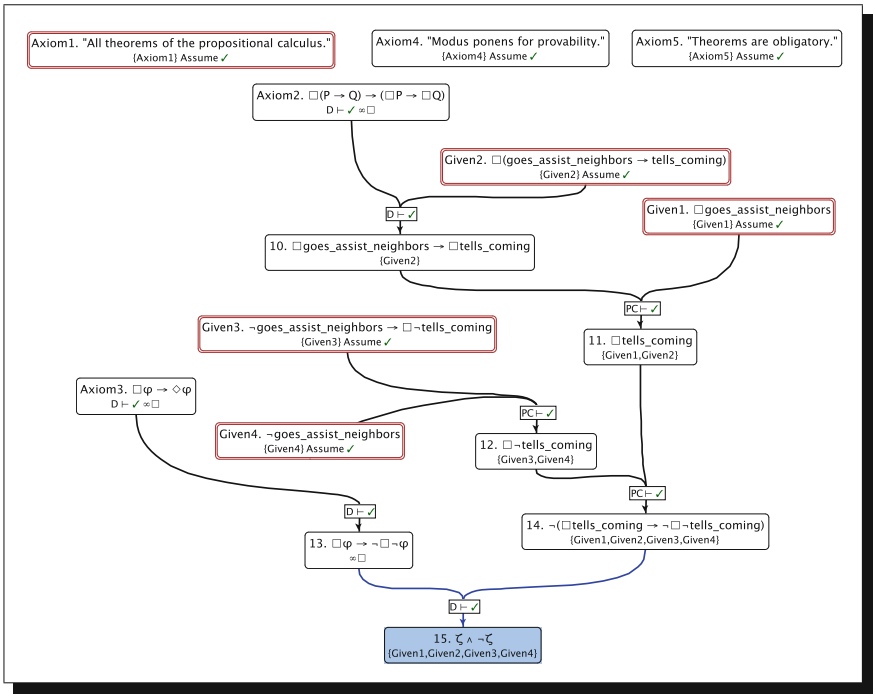


Fig. 4.1 Chisholm's paradox in standard deontic logic (proof in the Slate system of Bringsjord and Taylor)

plainly seen e.g. in Arkoudas et al. 2005). Of course, as cognoscenti will recall, it was Chisholm's (1963) Paradox (CP) that gave birth to deontic logic in earnest: we knew from the moment that his proof was published that simple use of the operators just given would lead immediately to inconsistency. Hence the kind of simple deontic logics laid out even over three decades after CP (e.g., in Chellas 1980), which unfortunately have found their way like a cancer into contemporary robot/machine ethics, are provably inconsistent (see the proof shown Fig. 4.1).

While RAIR-Lab robot-ethics engineering steers clear of Chisholm's Paradox, our logics provided thus far have been based on a dyadic operator **O**, and correspondingly dyadic versions of **P** and **F**; these logics are currently configured as an "ethical stack"; see Fig. 4.2. This figure gives a pictorial bird's-eye perspective of the high-level architecture of a system from the RAIR Lab that augments the DIARC (**D**istributed **I**ntegrated **A**ffect, **R**eflection and **C**ognition; see Schermerhorn et al. 2006) robotic platform with ethical competence. Ethical reasoning is implemented as a hierarchy of formal computational logics (including, most prominently, sub-deontic-logic systems) which the DIARC system can call upon when confronted with a situation that the hierarchical system believes is ethically charged. If this belief is triggered, our hierarchical ethical system then attacks the problem with

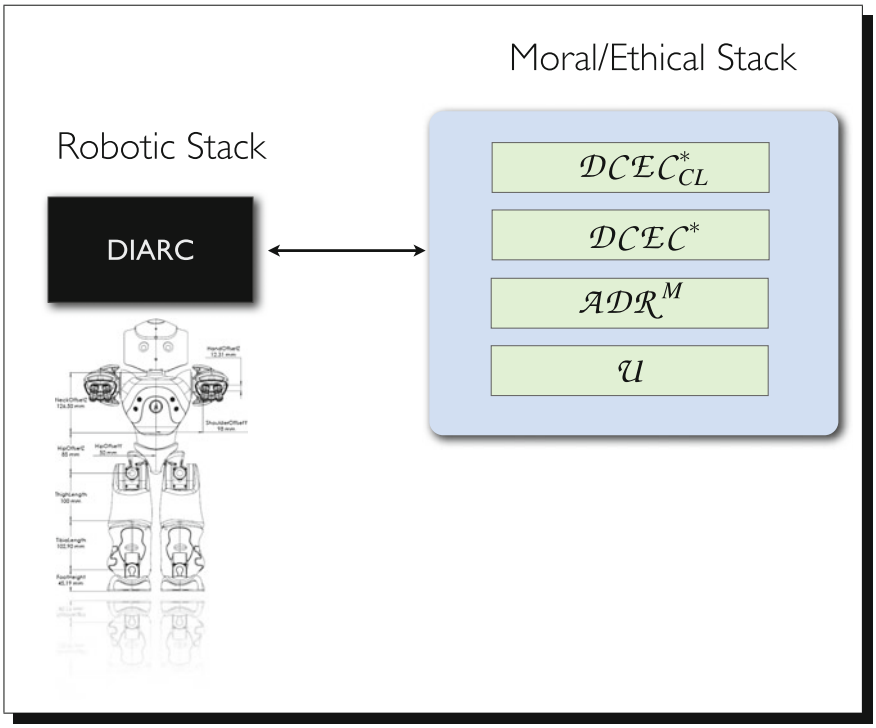


Fig. 4.2 Pictorial overview of non-$\mathscr{EH}$-based situation. The first layer, $\mathscr{U}$, is, as said in the main text, based on UIMA; the second layer on what we call *analogico-deductive reasoning* for ethics; the third on the "deontic cognitive event calculus" with an indirect indexical; and the fourth like the third except that the logic in question includes aspects of conditional logic (Robot schematic from Aldebaran Robotics' user manual for Nao. The RAIR Lab has a number of Aldebaran's impressive robots.)

increasing levels of sophistication until a solution is obtained, and then passes on the solution to DIARC. This approach, while satisfactory in the near-term, is ultimately inadequate for two reasons. One, the efficacy of this approach (and an expansion of the approach based on $\mathscr{EH}$), requires that implemented deontic logics have control at the operating-system level (an issue treated in detail in Govindarajulu and Bringsjord 2015). The second defect is that this hierarchy is based on the obsolete 19th-century hierarchy.

So what is this 19th-century hierarchy that underlies even much of my lab's contemporary work, and other work that partakes of underpinnings based on that which is forbidden, permissible, and obligatory (e.g., Arkin 2009)? We can set out the hierarchy by simply positing clusters of behaviors corresponding to the standard operators. For example, a creature that performs forbidden actions would fall into the cluster $\mathscr{F}$, whereas a creature whose performed actions meet obligations would fall into $\mathscr{O}$. Here then, given a self-explanatory way of picking out the set of morally

indifferent actions, is the obsolete hierarchy:

$$\mathscr{T} := \|\mathscr{F}|\mathscr{P} \wedge \neg\mathscr{O}|\mathscr{O}\|$$

It will be convenient if I next introduce, before passing to the hierarchy $\mathscr{EH}$, a mechanism, based on five quantifiers, for sub-categorizing a cluster of actions. Here are the five quantifiers in question, where of course the first and fifth will be familiar to all readers:

all: $\forall$
few: F
most: M
vast majority: V
at least one: $\exists$

The basic idea is straightforward. An agent that, within for instance the category $\mathscr{O}$, meets all its obligations, falls within the sub-category $\mathscr{O}^{\forall}$; an agent that meets only a few of its obligations falls within the sub-category $\mathscr{O}^{\mathrm{F}}$; and so on. (Inductive, rather than merely deductive, logics are needed to formalize the three non-standard quantifiers. The logic $\mathscr{L}_{\mathscr{EH}}$ is inductive, and below (Sect. 4.7) I say a few words about the proof-theoretic machinery needed for the quantifier M.) Here then is a basic picture of the new—but still fundamentally trichotomous—hierarchy $\mathscr{T}^{Q}$:

$$\begin{array}{ccccc|c|ccccc} & & \mathscr{F} & & & & & & \mathscr{O} & & \\ \forall & \mathrm{V} & \mathrm{M} & \mathrm{F} & \exists & \mathscr{P} \wedge \neg\mathscr{O} & \exists & \mathrm{F} & \mathrm{M} & \mathrm{V} & \forall \end{array}$$

At this point, two immediate confessions are in order, before proceeding. It will occur to the skeptical reader that the use of the existential quantifier in $\mathscr{T}^{Q}$ is peculiar. The reason is of course that in standard first-order logic $\vdash \forall x\phi(x) \rightarrow \exists x\phi(x)$. Accordingly, confession one: we don't here have a strict sub-hierarchy via quantification, such as is seen in the quantifier-based version of the Arithmetic Hierarchy (Davis et al. 1994). Options are available for fixing this quirk, but given space constraints I don't discuss them herein.[4] The second confession is that while there is no consequentialist fabric indicated by the bare bones of $\mathscr{T}^{Q}$, such a fabric is ultimately desirable to flesh out and exploit in some detail. The reason is that, with respect to their consequences, not all actions within the same sub-category are equivalent. In the real world, opportunity is an important factor in determining one's place in an ethical hierarchy. If Smith is locked in solitary confinement, the (leaving aside purely mental actions to ease exposition) range of obligations that bind him may be severely limited. Jones, a free man living in interaction with other humans, may in contrast be bound by numerous demanding obligations. If Jones manages to meet most of his obligations, and Smith does too, it would be counter-intuitive to classify Jones and Smith as both (over some fixed time interval) within $\mathscr{O}^{\mathrm{M}}$. In the present paper, which

[4] One option is of course to supplant $\exists$ with $\exists^{=1}$.

is intended to introduce $\mathscr{EH}$ and not to plumb its depths, I confess to ignoring this complication.

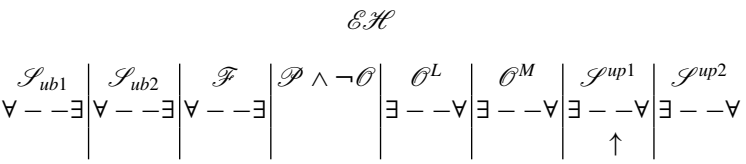
4.4 $\mathscr{EH}$, for the 21st Century

Why is the hierarchy $\mathscr{T}$ incomplete? Perhaps the quickest way to see that $\mathscr{T}$ is incomplete is to simply call to mind actions in the human sphere that are heroic or saintly, a category famously depicted by Urmson (1958). Each week the newspapers bring to us stories about people who perform actions that are good, but—to use Ladd (1957, p. 127) phrase—"not wrong not to do." In fact, such cases in the past are so numerous, and new ones are so easily imagined, that I will spend no time citing or concocting any at the moment.[5] (We shall consider below an excellent example from Scheutz and Arnold 2016 in the robot realm.) I suspect, in fact, that the reader has himself/herself performed some heroic acts of self-sacrifice: acts that were permissible, good, but not obligatory. In a world filled with poverty and disease, there are a lot of opportunities for heroic actions.

But as Chisholm (1982, p. 100) points out, supererogatory actions include not just those that are heroic, saintly, or self-sacrificial, but also actions that are courteous, polite, kind. Showing kindness to dogs, going out of one's way to pet them when coming upon them in the normal course of life; issuing compliments to those who clearly have invested much time in their appearance or a project; giving words of encouragement to colleagues who are under considerable pressure on the job—these actions compose a category of actions—called 'charitable' by Leibniz—that are supererogatory, but obviously not saintly or heroic. Accordingly, I divide supererogatory actions into two categories, the merely charitable ($\mathscr{S}^{up1}$), and the heroic or saintly ($\mathscr{S}^{up2}$). In addition, the flip side of these two categories exist on the "dark" side of $\mathscr{EH}$; that is, on the suberogatory side.[6] In addition, I roughly follow Leibniz and Grotius in sub-dividing duties or obligations into the less demanding legal ones that proscribe harm, and the more demanding general space of ethical obligations. Given that we preserve the five quantifiers used in $\mathscr{TQ}$, we have our new, comprehensive hierarchy (I include the up-arrow to mark the location of the military robots (currently uniquely) targeted by my lab):

[5] Anyone who has stood atop Pointe du Hoc and pondered the self-sacrifice of the Rangers who battled the Nazis there will confront the stark reality that supererogation was required to vanquish Hitler. Leibniz would say that the pursuit of such victory makes no sense if there is no God and no afterlife (for reasons explained in Youpa 2013)—but this claim is one left aside here. I note only that Leibniz thought it was easy enough to prove God's existence, so for him, an ethics that presupposed God's existence was in no way scientifically problematic.

[6] I don't have the space to consider the evil actions in question; Chisholm (1982) provides some examples. By the way, it seems to me very likely that robots capable of *sub*erogatory actions will prove to be quite useful in espionage, but this topic cannot be discussed the present short paper. Readers interested in this direction are advised to begin with (Clark 2008).



4.5 Why Robot Ethics Based on Laws Is Untenable

I provide two general reasons why machine/robot ethics based solely upon laws is inadequate, from the perspective of $\mathscr{EH}$. The first reason is perfectly straightforward and unsurprising: viz., that legal obligations are only a small proper subset of obligations. I may not be legally obligated to try to minister to a weeping colleague at work, but *ceteris paribus* I'm nonetheless morally obligated to do so. For Leibniz, and flowing therefrom into $\mathscr{EH}$, the *neminem laedere* principle that one shouldn't harm others is what generates obligations not to harm, and these are the "lowest" obligations (i.e., $\mathscr{O}^{L}$). One might say that Asimov's famous Three Laws of Robotics, discussed in (Bringsjord et al. 2006), fall within this sub-category of obligations; the trio is thus incomplete from the standpoint of $\mathscr{EH}$.[7]

A robot joining a human soldier on a mission might well fulfill all its legal obligations (relative, e.g., to "laws of war" and "laws of engagement") while at the same time by failing to meet a moral obligation to minister to a severely depressed soldier might endanger the very mission in question. And, since as I will explain in the next (Sect. 4.6) section, there are actions that are morally good, and indeed such that we would wish a robot on a mission to perform them, yet these actions aren't ethically obligatory.

The second reason why basing machine/robot-ethics on legal principles, at least in the military sphere, where (at least in the Occidental tradition) such principles are derived from, or at least directly reflective of, Just War Theory (JWT), is that extant law doesn't apply to cyberwarfare (Bringsjord and Licato 2015a, b).[8] In order to formulate new laws of cyberwarfare and cyberengagement, the human race is going to need to back up to deeper ethical principles, and then work out to a replacement of new laws of conflict. Absent the completion of this undertaking, which of course promises to be complicated and time-consuming, we are going to need to strive for machines/robots that are above $\mathscr{O}^{L}$ in $\mathscr{EH}$.

[7] The trio isn't only incomplete, but is just plain unacceptable. A robot medic or surgeon would routinely need to harm humans in order to save them. In saying this, I narrowly condemn Asimov's trio only. Ethically sophisticated contemporary engineers have worked out avenues by which robots can trade short-term harm for longer-term good; see e.g. Winfield et al. (2014).

[8] These papers thus provide a rigorous deductive case for a position at odds with the *Tallinn Manual on the International Law Applicable to Cyber Warfare* Schmitt 2013.

4.6 Why Robot Ethics Based on $\mathscr{T}$ Is Untenable

There are many reasons why the engineering of ethically correct robots needs to be based not on $\mathscr{T}$ or close variants thereof, but rather on $\mathscr{EH}$. In the interest of economy, I give only two here.

The first reason stems from the realities of human-robot interaction. Robots built to collaborate with humans but whose actions merely conform to what is obligatory, even if such robots fall into $\mathscr{O}^{\forall}$, would be highly problematic. I gave a case of this above, where a robot that fails to perform actions in $\mathscr{S}^{up1}$ in connection with a depressed soldier might endanger the mission the two are on.

The second reason why the engineering of moral robots needs to be based on $\mathscr{EH}$ is more interesting. The reason can be seen by considering a case described by Scheutz and Arnold (2016), in which a robot doing road repair with a jackhammer notices a child dart out to retrieve a bouncing ball, "a car speedily approaching and headed directly at her." Under the supposition that the car will not be able to stop before hitting the girl, and that the robot can move the young child to safety at the cost of losing its own life, what would have been the right kind of prior engineering here? Presumably the right sort of engineering would have been that which produced a robot that performs the supererogatory rescue of the girl.[9] Notice that even if one insists that the self-sacrificial rescue is an obligation for the robot, the fact remains that we must have robots able to consider that such actions, when performed by humans, are supererogatory. Hence we cannot dodge the need to engineer robots on the basis of the concepts that distinguish $\mathscr{EH}$ from $\mathscr{T}$.[10]

4.7 On the Logic $\mathscr{L}_{\mathscr{EH}}$

Hitherto, Bringsjord-led work on robot ethics has been unwaveringly logicist (e.g., see Govindarajulu and Bringsjord 2015); that's par for a course long set for human-level AI (e.g., see Bringsjord and Ferrucci 1998, Bringsjord 2008b) and its sister field computational cognitive modeling (e.g., see Bringsjord 2008a). Nothing in or about the hierarchy $\mathscr{EH}$ will change this trajectory. However, $\mathscr{EH}$ does reveal that the logics invented and implemented thus far in this trajectory (e.g., **deontic cognitive event calculi**, such as $\mathcal{D}^{\mathcal{e}}\mathcal{CEC}$) (Bringsjord and Govindarajulu 2013), are inadequate. For it can be seen that for instance the formal language and proof theory for $\mathcal{D}^{\mathcal{e}}\mathcal{CEC}$, shown in Fig. 4.3, contains no provision for the super/suberogatory.

I can offer only a few remarks about how the inadequacies in question are met in $\mathscr{L}_{\mathscr{EH}}$ (but see Footnote 1). In keeping with the compressed notation employed above, I suppress many of the elements that are in my lab's extant deontic logics.

---

[9] In the human sphere, such a rescue would clearly fall into $\mathscr{S}^{up2}$. For reasons pertaining to A- versus P-consciousness and the imaginary robot, I classify the rescue as a $\mathscr{S}^{up1}$ action.

[10] Thoroughgoing Kantians might resist $\mathscr{EH}$, and the robot ethics and robot-ethics engineering that seems to naturally flow from it (because Kantian/deontological ethical theories are obligation-myopic). This is an issue I'm prepared to address—but not in this short paper. Robot ethics as it relates to Kant should in my opinion begin with study of (Ganascia 2007), (Powers 2006).

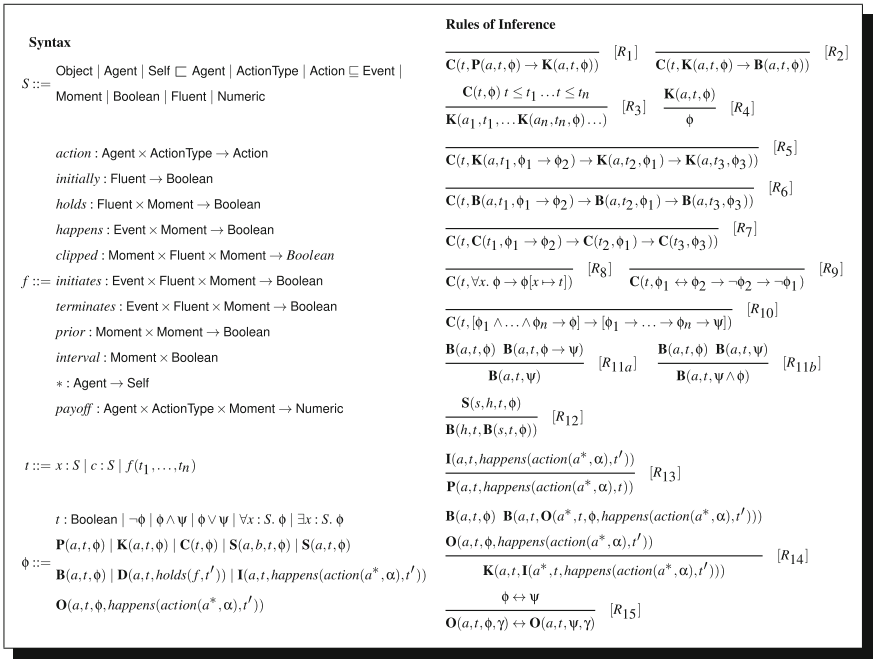


Fig. 4.3 $\mathscr{D}^{e}\mathscr{CEC}$ syntax/signature and rules of inference

There are obviously a host of formulae whose theoremhood constitute desiderata; that is (to give but a pair), the following must be provable (where $n \in \{1, 2\}$):

Theorem 4.1 $\mathbf{S}^{\mathbf{up}n}(\phi, a, \alpha) \rightarrow \neg\mathbf{O}(\phi, a, \alpha)$

Theorem 4.2 $\mathbf{S}^{\mathbf{up}n}(\phi, a, \alpha) \rightarrow \neg\mathbf{F}(\phi, a, \alpha)$

Secondly, $\mathscr{L}_{\mathscr{EH}}$ is an *in*ductive logic, not a deductive one. This must be the case, since, as we've noted, quantification isn't restricted to just the standard pair $\exists\forall$ of quantifiers in standard extensional n-order logic[11]: $\mathscr{EH}$ is based on three additional quantifiers. For example, while in standard natural deduction we have the inference schema

$$\frac{\forall x \phi}{\phi(\frac{x}{a})}$$

of universal elimination, how would such a thing work for the formula $\mathtt{M}x\phi$? The answer is that in $\mathscr{L}_{\mathscr{EH}}$ strength factors are assigned to formulae (in keeping with the 9 strength factors in Bringsjord et al. 2008), and every inference schema dictates

[11] Technically, what I've said here is incorrect, since some numerical quantifiers do work just fine with deduction. Example, from $\exists^{\geq k} x\phi$ we can deduce $\exists x\phi$.

the strength of inferred formulae from given formulae and the strength factors that they have. Standard inference schemata like universal elimination simply follow the "weakest link" principle.

Third, and I shall stop here, $\mathscr{L}_{\mathscr{EH}}$ not only includes the machinery of traditional third-order logic (in which relation symbols can be applied to relation symbols and the variables ranging over them), but allows for quantification over formulae themselves, which is what allows one to assert that a given agent a falls in a particular portion of $\mathscr{EH}$. So for example, one hopes that those charged with engineering robots for sensitive operations in the military and medical realms manage to engineer robots occupying the $\forall$ portion of the $\mathscr{O}$ portion of $\mathscr{EH}$; that is, one hopes that all robots r engineered by such people are such that $\mathscr{O}^{\forall}(r)$ holds, where

$$\mathscr{O}^{\forall}(r) \leftrightarrow \forall\phi\forall\alpha[\mathbf{O}(\phi, a, \alpha) \rightarrow happens(\alpha)]$$

4.8 A Note on Vacuous Quantification and $\mathscr{EH}$

It's quite important to note that some variants of our original question are trivial, because it's trivial to prove that an answer to them is correct.[12] (I'm indebted to Alexander Bringsjord for stimulating my coverage of this point.) I steered clear of considering for instance this trivial question:

(Q1) Can humans build some robots that will be more moral than some humans?

Given $\mathscr{EH}$, it's easy to prove that the correct answer to this question is in the affirmative. But no one should be aiming to build such morally mediocre robots; doing so is easy, and ultimately dangerous. Why is the answer to (Q1) "Yes"? The reasoning is simple in the context of $\mathscr{EH}$. Clearly, it's a brute empirical fact that there exist humans falling within the M portion of the $\mathscr{S}^{ub2}$ portion of $\mathscr{EH}$. (Any number of villains from human history fit the bill.) And yet, given what I have said about the SIFC quartet, it's logically impossible for any robot to place this low in $\mathscr{EH}$. Perhaps more pragmatically put, using techniques promoted by myself and others, it seems easy enough (given sufficient funding) to engineer a robot that falls within the V portion (or *at least* the M portion) of the $\mathscr{S}^{ub2}$ portion of $\mathscr{EH}$. But this seems insignificant within the overall landscape of robot ethics.

And now here is a variant of the original question that seems quite important:

(Q2) Can we engineer robots that meet *all* of their legal and moral obligations?

The answer to this one is Yes, and this is the question-answer pair that I see myself working toward demonstrating. But if what has been said above is correct, this is insufficient, because at least supererogatory actions should sometimes be performed as well.

[12] Those familiar with the quantifier-based version of the Arithmetic Hierarchy will wonder whether $\mathscr{EH}$ can likewise be built crisply via layered quantification. The answer, it seems to me, is Yes.

A final point: Obviously, I interpreted (Q) in such a way that it's logically equivalent to:

(Q′) Can humans build some robots that will be more moral than all humans?

which is in turn equivalent to:

(Q″) Can humans build some robots that will be more moral than the overall class (or capacity) of humans?

The answer to (Q′) and (Q″), again, for reasons given, is firmly in the negative.

4.9 Conclusion; Future Work

The hierarchy $\mathscr{EH}$ has only been sketched in the present, short paper; that will by now be clear to all readers.[13] The goal here has been to throw light on robot/machine ethics, revealing deep inadequacies (e.g., that of basing work on the incomplete and naïve tripartite hierarchy $\mathscr{T}$ now superseded (at least in my lab) by $\mathscr{EH}$). Obviously, then, future work must include full specification of $\mathscr{EH}$; and just as obviously, future work must include as well the concomitant specification, and indeed implementation, of $\mathscr{L}_{\mathscr{EH}}$.

Please allow me to conclude by saying that future work undertaken in response to $\mathscr{EH}$ shouldn't, in my opinion, be confined to my own work, and those in my laboratory. I strongly suggest that other researchers working in machine/robot ethics branch out, within their preferred methodology, to the super/suberogatory. For example, Bello (2005), Bello and Bringsjord (2013) could consider extending the reach of computational cognitive modeling to cover cognition associated with the parts of $\mathscr{EH}$ not present in $\mathscr{T}$. Pereira (2016) and colleagues could consider extending the reach of their powerful logic-programming paradigm to model the parts of morality reflected in $\mathscr{EH}$. And while Arkin's (2009) underpinnings have unfortunately hitherto been firmly in $\mathscr{T}$, and his focus has hitherto been also unfortunately firmly on laws, he should consider working from the broader underpinning of $\mathscr{EH}$ and its new sub-categories.[14]

[13] There are in fact two deep lacunae in what has been presented: two sub-parts of the hierarchy that are flat-out missing, one toward the endpoint of moral perfection, and one toward the endpoint of the diabolical. Both lacunae pertain to *intelligence*: it seems at least *prima facie* untenable to leave the level of intelligence of ethical agents out of systematic investigation of a continuum of ethical "grade".

[14] Within the robot-ethics project of which my logicist work is a part (see Acknowledgements), the empirical investigation of moral competence led by Malle can perhaps explore "norms" that cover not only what might naturally be classified within deontic logics as obligations, but also what conventional attitudes toward both levels 1 and 2 of supererogation in $\mathscr{EH}$. I wonder whether for example the everyday concept of blame, under exploration by Malle et al. (2012), extends to supererogation.

Acknowledgements Bringsjord is profoundly grateful for support provided by two grants from U.S. ONR to explore robot ethics, and to co-investigators M. Scheutz (PI, MURI; Tufts University), B. Malle (Co-PI, MURI; Brown University), M. Sei (Co-PI, MURI; RPI), and R. Sun (PI, Moral Dilemmas; RPI) for invaluable collaboration of the highest order.

References

Arkin R (2009) Governing lethal behavior in autonomous robots. Chapman and Hall/CRC, New York

Arkoudas K, Bringsjord S, Bello P (2005) Toward ethical robots via mechanized deontic logic. In: Machine ethics: papers from the AAAI fall symposium; FS–05–06. American Association for Artificial Intelligence, Menlo Park, CA, pp 17–23. http://www.aaai.org/Library/Symposia/Fall/fs05-06.php

Bello P (2005) Toward a logical framework for cognitive effects-based operations: some empirical and computational results. PhD thesis, Rensselaer Polytechnic Institute (RPI), Troy, NY

Bello P, Bringsjord S (2013) On how to build a moral machine. Topoi 32(2):251–266, http://kryten.mm.rpi.edu/Topoi.MachineEthics.finaldraft.pdf

Block N (1995) On a confusion about a function of consciousness. Behav Brain Sci 18:227–247

Bringsjord S (1992) What robots can and can't be. Kluwer, Dordrecht

Bringsjord S (1999) The zombie attack on the computational conception of mind. Philos Phenomenol Res 59(1):41–69

Bringsjord S (2007) Offer: one billion dollars for a conscious robot. If you're honest, you must decline. J Conscious Stud 14(7):28–43. http://kryten.mm.rpi.edu/jcsonebillion2.pdf

Bringsjord S (2008a) Declarative/logic-based cognitive modeling. In: Sun R (ed) The handbook of computational psychology. Cambridge University Press, Cambridge, pp 127–169. http://kryten.mm.rpi.edu/sb_lccm_ab-toc_031607.pdf

Bringsjord S (2008b) The logicist manifesto: at long last let logic-based ai become a field unto itself. J Appl Logic 6(4):502–525. http://kryten.mm.rpi.edu/SB_LAI_Manifesto_091808.pdf

Bringsjord S, Ferrucci D (1998) Logic and artificial intelligence: divorced, still married, separated? Minds Mach 8:273–308

Bringsjord S, Govindarajulu NS (2013) Toward a modern geography of minds, machines, and math. In: Müller VC (ed) Philosophy and theory of artificial intelligence, studies in applied philosophy, Epistemology and rational ethics, vol 5. Springer, New York, pp 151–165. doi:10.1007/978-3-642-31674-6_11, http://www.springerlink.com/content/hg712w4l23523xw5

Bringsjord S, Licato J (2015a) By *dis*analogy, cyberwarfare is utterly new. Philos Technol 28(3):339–358. http://kryten.mm.rpi.edu/SB_JL_cyberwarfare_disanalogy_DRIVER_final.pdf

Bringsjord S, Licato J (2015b) Crossbows, von Clauswitz, and the eternality of software shrouds: reply to christianson. Philos Technol 28(3):365–367. http://kryten.mm.rpi.edu/SB_JL_on_BC.pdf, The url here is to a preprint only

Bringsjord S, Taylor J (2012) The divine-command approach to robot ethics. In: Lin P, Bekey G, Abney K (eds) Robot ethics: the ethical and social implications of robotics. MIT Press, Cambridge, pp 85–108. http://kryten.mm.rpi.edu/Divine-Command_Roboethics_Bringsjord_Taylor.pdf

Bringsjord S, Zenzen M (2003) Superminds: people harness hypercomputation, and more. Kluwer Academic Publishers, Dordrecht

Bringsjord S, Arkoudas K, Bello P (2006) Toward a general logicist methodology for engineering ethically correct robots. IEEE Intell Syst 21(4):38–44. http://kryten.mm.rpi.edu/bringsjord_inference_robot_ethics_preprint.pdf

Bringsjord S, Kellett O, Shilliday A, Taylor J, van, Heuveln B, Yang Y, Baumes J, Ross K, (2006) A New Gödelian argument for hypercomputing minds based on the busy beaver problem. Appl Math Comput 176:516–530

Bringsjord S, Taylor J, Shilliday A, Clark M, Arkoudas K (2008) Slate: an argument-centered intelligent assistant to human reasoners. In: Grasso F, Green N, Kibble R, Reed C (eds) Proceedings of the 8th international workshop on computational models of natural argument (CMNA 8). University of Patras, Patras, Greece, pp 1–10. http://kryten.mm.rpi.edu/Bringsjord_etal_Slate_cmna_crc_061708.pdf

Chellas BF (1980) Modal logic: an introduction. Cambridge University Press, Cambridge

Chisholm R (1963) Contrary-to-duty imperatives and deontic logic. Analysis 24:33–36

Chisholm R (1982) Supererogation and offence: a conceptual scheme for ethics. In: Chisholm R (ed) Brentano and meinong studies. Humanities Press, Atlantic Highlands, pp 98–113

Chisholm R (1986) Brentano and intrinsic value. Cambridge University Press, Cambridge

Clark M (2008) Cognitive illusions and the lying machine. PhD thesis, Rensselaer Polytechnic Institute (RPI)

Davis M, Sigal R, Weyuker E (1994) Computability, complexity, and languages: fundamentals of theoretical computer science. Academic Press, New York

Ganascia JG (2007) Modeling ethical rules of lying with answer set programming. Ethics Inf Technol 9:39–47

Govindarajulu NS, Bringsjord S (2015) Ethical regulation of robots must be embedded in their operating systems. In: Trappl R (ed) A construction manual for robots' ethical systems: requirements, methods, implementations. Springer, Basel, pp 85–100. http://kryten.mm.rpi.edu/NSG_SB_Ethical_Robots_Op_Sys_0120141500.pdf

Ladd J (1957) The structure of a moral code. Harvard University Press, Cambridge

Malle BF, Guglielmo S, Monroe A (2012) Moral, cognitive, and social: the nature of blame. In: Forgas J, Fiedler K, Sedikides C (eds) Social thinking and interpersonal behavior. Psychology Press, Philadelphia, pp 313–331

Powers T (2006) Prospects for a Kantian machine. IEEE Intell Syst 21:4

Saptawijaya A, Pereira LM (2016) The potential of logic programming as a computational tool to model morality. In: Trappl R (ed) A construction manual for robots' ethical systems, Springer, Cham. http://centria.di.fct.unl.pt/~lmp/publications/online-papers/ofai_book.pdf

Schermerhorn P, Kramer J, Brick T, Anderson D, Dingler A, Scheutz M (2006) DIARC: A testbed for natural human-robot interactions. In: Proceedings of AAAI 2006 mobile robot workshop

Scheutz M, Arnold T (2016) Feats without heroes: norms, means, and ideal robotic action. Front Robot AI

Schmitt M (ed) (2013) Tallinn manual on the international law applicable to cyber warfare. Cambridge University Press, Cambridge, UK, This volume was first published in 2011. While M Schmitt is the General Editor, there were numerous contributors, falling under the phrase 'International Group of Experts at the Invitation of the NATO Cooperative Cyber Defence Centre of Excellence'

Suppes P (1972) Axiomatic set theory. Dover Publications, New York

Urmson JO (1958) Saints and heroes. In: Melden A (ed) Essays in moral philosophy. University of Washington Press, Seattle, pp 198–216

von Wright G (1951) Deontic logic. Mind 60:1–15

Winfield A, Blum C, Liu W (2014) Towards an ethical robot: internal models, consequences and ethical action selection. In: Mistry M, Leonardis A, Witkowski M, Melhuish C (eds) Advances in autonomous robotics systems. Lecture notes in computer science (LNCS), vol 8717. Springer, Cham, pp 85–96

Youpa A (2013) Leibniz's ethics. In: Zalta E (ed) The stanford encyclopedia of philosophy, the metaphysics research lab, center for the study of language and information. Stanford University. http://plato.stanford.edu/entries/leibniz-ethics

Chapter 5
Robots and Free Software

Wilhelm E.J. Klein

Abstract This article examines whether the arguments put forward by Free Software advocates in the context of computers also apply for robots. It summarises their key arguments and explores whether or not they appear transferable to robot cases. Doing so, it comes to the conclusion that, in the majority of cases, the reasons that may make the use of Free Software over proprietary software preferable in other technologies, equally apply in the case of robots.

Keywords Robots · Ethics · Free software · Freedom · Readability · Open source · Privacy · Ownership · Power

5.1 Introduction

There is little question about the impact the personal computer revolution, the advent of the internet, and the recent introduction of smaller, but no less connected gadgets, has had on everyone's lives. Naturally, this development has also brought about new legal, moral and political questions to be tackled. One of the earliest voice in this debate about the rights and wrongs of technology was that of the so called Free Software movement,[1] which, from the very beginning of the computer revolution, predicted multiple problems that would arise from the widespread use of proprietary software on personal, corporate and governmental computers. Their critique began with the early, clunky computers of the eighties, persisted through ever more compact machines and continues today where computers fit neatly in our pockets. Now, a new computing machine is on the horizon. One that is equipped with powerful sensors

[1] In the following, Free Software in the sense of "software that adheres to the principles of the free software movement", will always be capitalised to accentuate its difference from mere "freeware" or "shareware".

W.E.J. Klein (✉)
School of Creative Media, City University of Hong Kong, Hong Kong, China
e-mail: mail@wilhelmklein.net

M.I. Aldinhas Ferreira et al. (eds.), *A World with Robots*,
Intelligent Systems, Control and Automation: Science and Engineering 84,
DOI 10.1007/978-3-319-46667-5_5

and motors and probably soon able to walk around. Accordingly, it may be worth examining whether the arguments that have been made with regards to our limb-less computers also apply to the software-operated machines poised to be the next big thing in our daily lives–robots.

As the essay attempts to provide a first examination in this direction, it is important to note that it is not an in-depth examination of the ethical foundations of the Free Software movement,[2] not a legal analysis and also not an unequivocal endorsement.[3] Rather, it is intended as a primer, an experimental application of certain arguments to test their possible viability for the efforts of robot ethics and potentially to spur further investigation into the matters they outline.

To my knowledge, the closest someone has come to a similar analysis was M. Ryan Calo in his *Open Robotics* report, where he was trying to find an optimal balance, or trade-off, between the innovation-bolstering effect of "openness" and the better legal protection/safety of "closeness" for robots, Calo (2010a). In that sense, however, his understanding of the concept of "openness" rather resembles that that of the "Open Source" community. Although clearly growing out of certain libertarian beginnings that were shared with the Free Software movement, today, in practice, Open Source advocates appear to care less about the ethical background of Open Source and much more about it's successful work methodology, Berry (2004), Chopra and Dexter (2005). Free Software, in contrast, appears to put a lot more weight on the moral dimensions of software design, use and regulation, which are generally framed in the form of principled, quasi-deontological notions of freedoms that a piece of software has to protect and adhere to, Berry (2004). In the words of Free Software advocates themselves, Open Source and Free Software "stand for views based on fundamentally different values. […] For the free software movement, free software is an ethical imperative, essential respect for the users' freedom. By contrast, the philosophy of open source considers issues in terms of how to make software "better"—in a practical sense only. […] Most discussion of "open source" pays no attention to right and wrong, only to popularity and success;" Stallman (2015b). Although slightly charged, and maybe a bit biased,[4] Stallman outlines exactly the aspect of "openness" Calo focused on, the (methodological) practicality. This essay, in contrast, sets out to reflect on robotics from the, arguably more morally charged point of view of Free Software.

[2]For discussions of this kind see, for example, Jesiek (2003), McInerney (2009), Berry (2004).

[3]Clearly, there is still considerable room for debate and potential for improvement on the Free Software principles and licenses. For a very recent new development connected to the use of GPL licensed software in military technology, see Schroeder (2015).

[4]Given the fact that the hacker/programmer landscape is quite diverse, Coleman and Golub (2008), and the evidence that at least some contributors to open source are also driven by what could be called moral reasons, Bonaccorsi and Rossi (2006).

5.2 Free Software, PCs and Robots

Until the early 1970s, Free Software did not really exist as a concept itself as, by default, software was written, used and shared in the way now associated with Free Software researchers, programmers and organisations would freely exchanged software, sanctioned by hardware manufacturers who profited from this modus operandi as their products were useless without software to run on them. At the beginning of the 1970s this picture began to change due to a couple of partly independent, partly interconnected reasons. Firstly, software had become more complex and costlier to produce, which lead some to realise there were opportunities to generate or increase profits by decreasing the readability and shareability of software. Secondly, hardware manufacturers became weary of providing machines that had service contracts attached to ensure interoperability at an extra cost for the company Fisher et al. (1983), Stallman (1999). Thus, in collusion with certain hardware manufacturers, some software creators began to sell their software in a form that prevented others from sharing or changing it as they used to be able to–the advent of proprietary software.

In 1983 Richard Stallman, one of the earliest computing pioneers and hacker in the original sense of the word,[5] announced the GNU project as a personal reaction to these developments. In his initial announcement, he outlined his rejection of "non-disclosure agreements" and software licenses, arguing that these would not be compatible with his personal principles, Stallman (1983). Since then, his stated principles have served as the moral foundations of the Free Software movement and have been fleshed out and refined in the current form of four essential freedoms:

- "The freedom to run software as you wish, for any purpose (freedom 0).
- The freedom to study how the software works, and change it so it does your computing as you wish (freedom 1). Access to the source code is a precondition for this.
- The freedom to redistribute copies so you can help your neighbour (freedom 2).
- The freedom to distribute copies of your modified versions to others (freedom 3). By doing this you can give the whole community a chance to benefit from your changes. Access to the source code is a precondition for this." Free Software Foundation (2015b).

The principles are built on the premise that computers are universal machines. It will do anything you want it to do, as long as you provide proper instructions. These instructions, and the format they come in lie at the heart of the matter. They can either be open or free, which means that their source code, the human-readable code written in a programming language and accompanied by additional instructions and comments is provided. Or, the instructions can come as lower-level code only, devoid of commentary and programming language, decipherable only by computers (raw ones and zeros) but incomprehensible for a human users. If that is the case

[5] In the sense of Steven Levy's *Hackers: Heroes of the Computer Revolution*, (1984), not the sense of a malicious attacker, as this term is often misused today.

and the source code is not provided, a user can only observe results and guess the functions that lie beneath. One is effectively working with a software black box.[6] It is also impossible to modify the program, use differently than predetermined by its producer or connect it to third party software.

The ability to maintain full control over a machine, to read and modify the instructions, constitutes the freedom in Free Software. If the instructions, first and foremost, remain under the control of a third party, leaving the user only partially or in no control at all over the machine, the software is considered non-Free as it fails to protect and adhere to the user's freedom. Phrased not in terms of freedoms but preferences or interests, one could also say that software primarily either adheres to the user's interests or it follows the interests of whoever maintains control. Or as Richard Stallman puts it: "With Free Software, the users control the program, both individually and collectively […]. With proprietary Software, the program controls the users, and some other entity (the developer or "owner") controls the program", Stallman (2013).

This "control over the user" may sound quite Orwellian, but in general, these power dynamics are not part of some sinister plan by companies like Microsoft, Apple or Google. Rather, they are owed to the fact that products tend to inherit the quintessential values of their producers, which, in most cases are for-profit organisations, Brey (2010). As such, they are subject to the same profit maximisation imperative as any other corporation, and thus often sacrifice the rights and freedoms of their users in order to secure their businesses interests, Zittrain (2009), McChesney (2013). Consequently, their products end up being oppressive, restrictive etc. as well. Not because of malicious intent, but because it is necessary, or optimal, to protect or maximise profits.

As mentioned above, the Free Software principles are generally regarded as a deontological framework, based on Richard Stallman's understanding of freedom and liberty (which ought not to be infringed upon). At the same time however, one can approach the framework from a consequentialist or virtue ethics point of view, as it always emphasises the negative consequences that may arise from the use of non-Free software. The application cases that will be outlined below, for example, are all taken from the various publications of the Free Software movement and are all about particular (potential) negative *consequences* in areas like education and empowerment, personal freedoms and capabilities and so on, and additionally criticise the potential harm to the *virtues* and *values* of a individuals and society.

So far, these arguments have mostly been raised in the context of classic computing machines like personal computers etc. In the following, this essay will test whether their arguments remain valid when applied to robots.

[6] Admittedly, a case could be made that, when roboticists work with open ROS libraries, for example, they are working with black boxes as well, as they often do not bother to dig deeper into the code. This does not challenge the most important part of the argument however, as it focuses on capability in principle, not actual use in practice. The point is not, that everyone really is or should be auditing the code, but that everyone could do so in principle.

5.2.1 Education and Empowerment

Only with Free Software, advocates argue, can schools and other educational institutes stay true to their essential purposes, which is to "disseminate human knowledge and to prepare students to be good members of their community", (2015a). As the primary shapers of society, schools are considered to have the moral obligation to teach exclusively Free Software as doing otherwise would result in teaching students to be dependent and subservient.[7] In contrast, making use of, and teaching Free Software would teach students "the habit of cooperating, helping each other". The use of proprietary software forces teachers to punish students "who are good enough at heart to share software" and "those curious enough to want to change it", Stallman (2013). The use of Free Software on the other hand, would encourage such behaviour.

This argument appears relatively easy to transfer. Already today, many successful classes teaching robotics to students in schools around the world work with open hardware and software–for example, the popular and open Arduino platform, Kuchment (2012). Indeed, the principles outline above appear to make particular sense in the case of robots. If you want to teach a student more than a black-box, "press this, get that", model of robotics, one has to be able to reveal what is going on beneath the hood. With regards to dependence and subservience, one can easily imagine, if not foresee that future generations will be living with robots, police robots, medical robots etc. in their daily lives. In any of these cases, it appears that the ability to approach and engage with well functioning, but especially with compromised, possibly virus-infected robots in an informed, not a completely speculative black-box manner may be of considerable advantage. One may not be able to read the full code of a rouge robot one encounters, but it may help to have a good, basic understanding of it's inner workings—which is only possible if, at some point in one's education, one had the chance to look at the actual functions of such machines.

5.2.2 Restriction and Locking in

Proprietary software usually comes with various kinds of *digital restrictions management* (*DRM*), essentially an artificial crippling of software in order to control what one can and cannot do with it. Firstly, this prevents a user from sharing the software with others, including his/her friends, family, and often even among his/her own computers. Secondly, it keeps the user from manipulating the software itself. In addition, it also functions as mechanism to lock users into particular software environments. From a corporate perspective, this ensures that users continue to use, and pay for software produced by them or at least by one of their sanctioned, con-

[7] An argument reminiscent of Amaryta Sen and Martha Nussbaum's capabilities approaches, which also placed much emphasis on notions of freedom and self-determination, Sen (1999), Nussbaum (2011).

tractual partners. As Eben Moglen, chairman of the Software Freedom Law Center, puts it: "[t]he purpose of locked down operating systems and the embedded software stores is not to enable people to make informed software choices, but to conceal alternatives, shape consumers and create proprietary service platforms they can't get away from", Moglen (2012). The Free Software GNU/Linux operating system platforms on the other hand provide the possibility to run whatever software one pleases, from whatever provider and at whatever price. The important difference is that the user maintains control and his or her choices and decisions are not restricted or pre-determined by a third party.

Relating this to robots, one can easily imagine a situation where branded robots come with locked down operating systems and in-house app stores, allowing installation only of pre-approved programs. In such cases, even though there might be cheaper, more efficient, or physically safer alternatives available in principle (e.g. in the form of Free Software), one would not be able to install them. The consequences could range from simple inconvenience and the loss of some money to the literal saving of lives. Imagine, for example, a company had developed a superior algorithm for collision prevention between robots (self-driving cars for example). What if this program was protected by DRM and only to be used in branded cars? How many lives might this cost? From an ethical point of view, a case could be made that every robot should be allowed to run this program if they are physically able to do so, prioritising the lives and safety of people over the protection of corporate profits.

5.2.3 Censorship

Thanks to the control creators of proprietary software maintain over their products, they are also able to censor, among other things, for content or third party software. In August 2012, for example, artist and developer Josh Begley created an iPhone app that would track and display drone strikes around the world committed by the U.S. military. Apple rejected the application multiple times, first arguing that it was not "useful or entertaining enough", then over unsubstantiated issues with the logo and finally because they deemed it "objectionable and crude", Wingfield (2012), clearly (albeit not officially) taking a political stance on the issue. There have also been accusations that Apple was, for a brief time, actively rejecting applications aimed at competitors' products. Apple quickly denied such claims and called them simple and innocent mistakes, but clearly there remains a certain suspicion, given how useful those mistakes appeared to be, Pierce (2015). Apple is of course not the only perpetrator. Microsoft, Google, AOL and many others, for example, decided to cooperate with the Chinese government its censorship programs so that they would not forego the immense profits expected from the growing Chinese market, Human Rights Watch (2006). Of course there are many more. One only has to type "[proprietary software company name] + censorship" into any search engine to find dozens of cases of complicity or censorship on the part of companies in order to protect their business interests.

With regards to censorship as the result of the protection of interests, it is easy to imagine, for example, a household robot allowed to perform only certain tasks. As a potential universal machine, if run on Free Software, one could teach a robot anything its physique permits. With a robot run on proprietary software, one can conceive a producer restricting its abilities based on various criteria. For example, the robot could be prohibited from creating a burger sauce resembling that of a McDonald's Big Mac, or from using its tools to reproduce itself, or to construct a third party product–or it could be prevented from closing the blinds if the government wants to see through your windows. To draw a parallel with the *great firewall of China*, one could, for example, imagine self-driving cars tainting their windows whenever they drive through prohibited (due to environmental damage, social uprising etc.) zones. Literally anything that would infringe on its producer's interest or the interest of entities the producer has third party contracts, or is in collusion with, could potentially be banned from installation and use. Whether you want your robot to speak with, say, Pikachu's voice, perform maintenance on your branded car, or point out which of the night sky stars is a NSA spy satellite, with Free Software, in principle, nothing stands in your way. With proprietary software, all of this could potentially be blocked, adding whole new dimensions to the "the software controls the user" conundrum as mentioned above.

5.2.4 Governance

As advocates argue, Free Software is necessary to maintain "computational sovereignty" on a governmental level: "Most government activities now depend on computing, and its control over those activities depends on its control over that computing. Losing this control in an agency whose mission is critical undermines national security", Stallman (2015a). Aside from this, there is also another level to the interplay of governments and the use of software. Many countries have track records of employing proprietary software for questionable tasks. This may range from general censorship, to surveillance, privacy infringement and even the active planting of compromising material, Burton (2015). Although such actions are mostly associated with dictatorships and other extremist regimes (see Deibert et al. 2008, 2010; MacKinnon 2012) they are equally employed by countries with much better reputation. The German government, for example, ordered the production of an unconstitutional piece of malicious software dubbed "Staatstrojaner" (Governmental Trojan Horse) in order to spy on suspects of various kinds, Solon (2011). Equally, there is very good reasons to believe that Stuxnet, one of the most sophisticated pieces of (non-Free) malicious software ever written, has found its beginnings at the National Security Agency of the United States, Appelbaum and Poitras (2013).

Particularly the first argument appears to be of outstanding significance for robots. If one imagines a government that has fully embraced the use of robots and is widely employing them for, say, postal services, civil engineering, fire fighting, police services as well as their military, the importance of computational sovereignty simply

cannot be understated. If such robots are operated by proprietary software, the government is only nominally in control. In reality, the companies who manufactured them maintain control, and could potentially hold an entire nation hostage–a case of moral hazard of epic proportions.[8] The second dimension (imagine an actual robotic State Trojan Horse) also appears to be exacerbated when applied to the case of robots. If governments are already willing to make use of malicious proprietary software to further their (often questionable) interests, there is no reason to doubt they would not also do the same with robots. With Free Software, at least in principle, a citizenship (and the watchdog organisations it may possess) would be able to monitor the functions and actions of governmental robots.

5.2.5 Insecurity

Any program of reasonable complexity will inevitably possess a certain amount of software bugs and potential security holes. This is simply because software engineers are (still) humans and occasionally make mistakes. This is true for both proprietary and Free Software. The important difference lies in the relative ability to identify and address these problems. In the case of Free Software, users and third party developers are able to discover and report these bugs and contribute to a solution to a problem–as happened, for example, with the so called "heartbleed" bug, a critical vulnerability in the popular OpenSSL cryptographic software library widely used to secure communication on the internet, Sullivan (2014). In the case of proprietary software, nobody besides the original software manufacturer can audit the code and/or implement fixes, should a problem be discovered. For this reason, as John Sullivan, executive director at the Free Software Foundation states, "[…] it is impossible to have a true chain of trust. Everyone is helpless until (the software company) decides to act", Sullivan (2014). These insecurities may have various consequences. For one, they may enable malicious attackers to break into sensitive computer networks to steal data. In 2013, for example, a large botnet (a network of infected and remote controlled computers) was discovered on point-of-sale terminals (the small computers used in stores, restaurants and other businesses to process credit and debit cards) operating on proprietary software and stealing the information of thousands of customers, Goodin (2013). More direct, and potentially lethal consequences might result from the widespread use of (insecure) proprietary software in medical equipment. A study in 2014 found that vitally important equipment such as drug infusion pumps, implanted defibrillators, X-ray machines, medical refrigerators and many more appliances were shockingly easy to reach, access and manipulate via the internet or internal hospital networks, potentially jeopardizing the lives of everyone depending on these

[8]To mention an example that renders this a real possibility, consider that already today, voting machines in the United States are run on proprietary software. Already today, there is no way to verify their proper functioning and to rule out systematic voter fraud. They remain under the full control of corporations and nobody else, Wasserman (2016).

devices, Zetter (2014). Some insecurities of this kind are unintentional programmers' mistakes, but many more are actually intended to be deliberate back doors, used to protect the interests of manufacturers (and their partners). They allow companies to gain direct access and control over the software, enabling them, for example, to read, modify or delete content and third party software. Apple, Microsoft, Google and Amazon all implement these back doors in their products Beaumont (2008), Keizer (2011a, b), Pogue (2009). These back doors are no more secure than any other piece of software and provide the perfect entry points for any malicious attacker. In other words, the manufacturers place much higher value on the protection of their own (business) interests than on the protection of the users of their software. To make this more accessible, imagine, for example, a piece of critical medical software as a literal bunker designed to provide the best possible protection from unauthorised access. For this reason alone, even if one were to dismiss all other arguments, the fact that no Free Software would ever come with an intentional back door serves as argument enough to prove its superiority in terms of security.

This does not seem to change when robots enter the equation. Programs run on robots will still have plenty of software bugs and corporations will still have many incentives to implement their back doors. As bad as it can be when someone uses the vulnerability of your computer or phone to steal information or recruit it for a botnet, imagine how potential damage could be amplified by a robots additional capabilities. A computer just sists there on your desk. A robot could potentially get up and deal very immediate, physical harm. What if, for example, someone used vulnerabilities in the operating software of governmental robots? In that case, what is usually considered science fiction could become reality, as someone could actually, in the not too far future, conceivably assemble his or her very own robot army by taking over a country's robot police force (if not military force). As mentioned above, already today some medical equipment suffers from critical vulnerabilities. Some of this equipment are already robots. One could argue that it is only a matter of luck that we have not seen a deadly attack on such medical machines yet. In a more private setting too, on can imagine problems arising from proprietary software-induced vulnerabilities. Imagine, for example, a multi-purpose household robot tasked with keeping watch over your toddler. What if that robot suddenly fell victim to a critical vulnerability and allowed an attacker to issue certain commands or to execute full remote control? Clearly, already in principle, robots need to be as secure as any other computing machine. But because of their additional abilities, every argument in this direction is amplified significantly, providing even stronger reasons to opt for Free Software whenever possible.

5.2.6 Privacy, Surveillance, and State Level Spying

Strongly related to the sections above, the use of non-Free software also affects matters of privacy and information sovereignty. On a business level for example, companies have many incentives to collect all kinds of information such as geo-

location, personal interests, consumer behaviour or social relationships, either to improve their own services, or to sell the data to third party businesses–always, of course, with the aim to increase their profits. Sometimes such practices are their very business model, as, for example, most social networks are all about the collection of data and pedling of products. One may argue whether such "free models" where the user appears to be the customer but, in reality, is the very product being sold, are ethical or not, but at least a veneer of consent is maintained when users are asked to accept the terms of conditions. The problem is, however, that most of these services are based on non-Free software, which means that users cannot know what else, aside from their consented self-commodification, these networks are doing with their data. On a governmental level, many states appear to have strong interests in collecting as much information as possible as well, both of their own and foreign citizens, corporations and other organisations. This is justified by its supposed usefulness to ensure national security and facilitate the fight terrorism.[9] As the Edward Snowden documents have proven, such efforts are often not just unconstitutional, but defying basic human rights, Savage and Weisman (2015). For decades Free Software advocates and security experts have warned that non-Free software would be an enabler of Orwellian governmental surveillance. Edward Snowden finally validated and unquestionably verified these claims. As predicted, proprietary software has served as the prime enabler of the sweeping, global mass surveillance of virtually everything happening on the internet and networks around the world. As NSA, GCHQ and other governmental agencies told companies to either collaborate or face legal consequences, only very few companies chose the former. Lavabit, for example, the email service used by Edward Snowden decided to shut down its business in order to protect its users, Ingraham (2013). The overwhelming majority followed their corporate imperatives to protect profits and provided, for example, early access to critical vulnerabilities, Gallagher (2013), or access to their in-house back doors, Hathaway (2013). Yet the problem does not stop at the level of citizen privacy. As the Snowden documents also revealed, governments actively engaged in industrial espionage to protect or push the competitive edge of domestic companies Greenwald (2014). All of this could only stay hidden because nobody, aside from the collaborating manufacturers of proprietary software, could examine and audit the code. One would think that such revelations would lead to public pressure and cause governments to rethink their practices. But considering, for example, the latest proposed legislature in the United Kingdom (mandatory encryption back doors, suppression of free speech etc., Masnick 2015), government appear to have vested interest in maintaining the status quo.

In terms of relevance, these issues may be the most significant for our current level of robotics and its societal impact. While complex robots with the potential to cause physical harm may still be a bit further in the future, many simpler robots are already equipped with all the necessary sensors to greatly aggravate the situation of illegal spying and surveillance Calo (2010b), Denning et al. (2009). Who knows what the

[9] Often, such governmental reasons are not quite comprehensible as they appear absolutely useless, Schneier (2015a, b).

next generation of vacuum robot, or childrens' toy sees[10] and hears aside from what is relevant to its original task. Projecting further into the future, it is conceivable that, like our smartphones and laptops today, potent robots will soon be accompanying us everywhere we go, and surround us at all times. However, unlike today's phones, which spend most of their time in dark pockets and with only limited opportunities to collect data, robots will very likely be able to use their sophisticated array of sensors much more freely and liberally. Consequently, they will be able collect a lot more information than any other machine can at present. Through the same mechanism that allow the NSA to maintain effective control over Google, Facebook etc., any robot running proprietary software could be forced to become part of a governmental spy programme and there would be no way to tell.[11] It appears that only Free Software operated robots would allow you to look inside and make sure it only does what it is supposed to do and nothing else.

5.3 Future Prospects

If Bill Gates' predictions prove true, and robots were 2007 where personal computers were 30 years prior, Gates (2007), then we are only a couple of years away from a transformation of society that could rival the industrial revolution or the one resulting from the computer revolution. This means, just as there was at the onset of the computer revolution, there will be a brief window of opportunity to get things right from the very beginning. Free Software advocates would probably argue that we missed this opportunity when we allowed software to be transformed and artificially crippled to become money-making, proprietary products, effectively setting up the preconditions for all the problems outlined above. They would probably say that we should not make the same mistake again.

Considering the arguments, I believe a case can be made that the use of Free Software is just as important for robots as it is for general computers. They too are potentially universal machines capable of performing any task you assign to them. They too, if run on non-Free software, are black boxes with potential, hidden dangers. But they come with all kinds of other capabilities exceeding those of regular computers. They possess audio-visual sensors better than any human eye or ear; they have arms legs and motors more powerful than any human limb and, unfortunately, they may even come with deadlier weaponry than any human could ever carry. One can argue about certain nuances of the Free Software movement, but considering

[10] Already today, you can find a first, primitive example of this in private households. Mattel's "Hello Barbie" already comes equipped with microphones etc., which it passes on to it's manufacturer and which can potentially be hijacked, Gibbs (2015).

[11] One may want to argue that, surely, no organisation would go to such lengths and spend the amount of money needed for such operations. I recommend watching the very illuminating and shocking lecture "To Protect And Infect, Part 2" delivered by Jacob Applebaum at the Chaos Communication Congress 2013, Applebaum (2013), to see how this argument vanishes in the face of the simply incredible lengths the NSA is willing and able to go in order to achieve their goals.

how their predictions have come true in the case of general computers, it may be worth to at least considering their ideas and principles in the case of machines where the stakes are potentially so much higher.

References

Appelbaum J, Poitras L (2013) Snowden interview: NSA and the Germans 'in bed together'. Der Spiegel. http://www.spiegel.de/international/world/edward-snowden-accuses-germany-of-aiding-nsa-in-spying-efforts-a-909847.html

Applebaum J (2013) To protect and infect, Part 2

Beaumont BC (2008) Apple's jobs confirms iphone 'kill switch'. http://www.telegraph.co.uk/technology/3358134/Apples-Jobs-confirms-iPhone-kill-switch.html

Berry DM (2004) The contestation of code. Crit Discourse Stud 1(1):65–89. doi:10.1080/1740590041000167452 4

Bonaccorsi A, Rossi C (2006) Comparing motivations of individual programmers and firms to take part in the open source movement: From community to business. Knowl Technol Policy 18(4):40–64. doi:10.1007/s12130-006-1003-9, http://link.springer.com/article/10.1007/s12130-006-1003-9

Brey P (2010) Values in technology and disclosive computer ethics. In: Floridi L (ed) Handbook of information and computer ethics. Cambridge University Press, Cambridge, pp 41–58

Burton G (2015) Did hacking Team design software that could plant child porn on suspects' PCs? Computing. http://www.computing.co.uk/ctg/news/2416521/did-hacking-team-sell-software-to-plant-child-porn-on-suspects-pcs

Calo R (2010a) Open robotics. SSRN Scholarly Paper ID 1706293, Social Science Research Network, Rochester, NY

Calo R (2010b) Robots and privacy. SSRN Scholarly Paper ID 1599189, Social Science Research Network, Rochester, NY

Chopra S, Dexter S (2005) The political economy of open source software. Int J Technol Knowl Soc 1(7):127–134

Coleman EG, Golub A (2008) Hacker practice—Moral genres and the cultural articulation of liberalism. Anthropol Theory 8(3):255–277. doi:10.1177/1463499608093814, http://ant.sagepub.com/content/8/3/255

Deibert R, Palfrey J, Rohozinski R, Zittrain JL, Gross Stein J (2008) Access denied: The practice and policy of global internet filtering. MIT Press

Deibert R, Palfrey J, Rohozinski R, Zittrain JL, Gross Stein J (2010) Access controlled: the shaping of power, rights, and rule in cyberspace. MIT Press

Denning T, Matuszek C, Koscher K, Smith JR, Kohno T (2009) A spotlight on security and privacy risks with future household robots: attacks and lessons. In: Proceedings of the 11th international conference on ubiquitous computing. ACM, New York, pp 105–114

Fisher FM, McKie JW, Mancke RB (1983) IBM and the U.S. data processing industry: an economic history. Praeger

FSF (2015a) Free software and education—How does free software relate to education. https://www.gnu.org/education/

FSF (2015b) What is free software? The free software definition. www.gnu.org/philosophy/free-sw.html

Gallagher S (2013) NSA gets early access to zero-day data from Microsoft, others. Ars Technica. http://arstechnica.com/security/2013/06/nsa-gets-early-access-to-zero-day-data-from-microsoft-others/

Gates B (2007) A robot in every home. Sci Am 296(1):58–65

Gibbs S (2015) Hackers can hijack Wi-Fi hello barbie to spy on your children. The Guardian. http://www.theguardian.com/technology/2015/nov/26/hackers-can-hijack-wi-fi-hello-barbie-to-spy-on-your-children

Goodin D (2013) Credit card fraud comes of age with advances in point-of-sale botnets. Ars Technica. http://arstechnica.com/security/2013/12/credit-card-fraud-comes-of-age-with-first-known-point-of-sale-botnet/

Greenwald G (2014) The U.S. government's secret plans to spy for american corporations. The Intercept. https://firstlook.org/theintercept/2014/09/05/us-governments-plans-use-economic-espionage-benefit-american-corporations/

Hathaway J (2013) The NSA has nearly complete backdoor access to Apple's iPhone. The Daily Dot. http://www.dailydot.com/politics/nsa-backdoor-iphone-access-camera-mic-appelbaum/

HRW (2006) "Race to the bottom" Corporate complicity in chinese internet censorship. Technical report, Human Rights Watch. https://www.hrw.org/reports/2006/china0806/

Ingraham N (2013) Lavabit founder closed his secure email service to 'protect the privacy' of its users. The Verge. https://www.theverge.com/2013/8/10/4608664/lavabit-founder-closed-his-secure-email-service-to-protect-the

Jesiek B (2003) Democratizing software: open source, the hacker ethic, and beyond. First Monday 8(10). http://128.248.156.56/ojs/index.php/fm/article/view/1082

Keizer G (2011a) Google throws 'kill switch' on Android phones. Computerworld. http://www.computerworld.com/article/2506557/security0/google-throws--kill-switch--on-android-phones.html

Keizer G (2011b) Microsoft: we can remotely delete Windows 8 apps. Computerworld. http://www.computerworld.com/article/2500036/desktop-apps/microsoft--we-can-remotely-delete-windows-8-apps.html

Kuchment A (2012) Hot bots: how Arduino teaches kids the science behind modern gizmos. Sci Am. http://blogs.scientificamerican.com/budding-scientist/arduino-101-build-halloween-monsters-blinking-clothing-and-more/

Levy S (1984) Hackers: heroes of the computer revolution. Doubleday, New York

MacKinnon R (2012) Consent of the networked: the worldwide struggle for internet freedom. Basic books

Masnick M (2015) UK Government goes full orwell: snooper's charter, Encryption backdoors, Free speech suppression|Techdirt. Techdirt. https://www.techdirt.com/articles/20150528/07001931137/uk-government-goes-full-orwell-snoopers-charter-encryption-backdoors-free-speech-suppression.shtml

McChesney RW (2013) Digital disconnect: how capitalism is turning the internet against democracy. New Press

McInerney PB (2009) Technology movements and the politics of free/open source software. Sci Technol Human Values 34(2):206–233. doi:10.1177/0162243907309852, http://sth.sagepub.com/content/34/2/206

Moglen E (2012) Eben Moglen explains freedom and free software. Slashdot. http://yro.slashdot.org/story/12/10/18/1759218/eben-moglen-explains-freedom-and-free-software-in-two-video-interviews

Nussbaum MC (2011) Creating capabilities. Harvard University Press

Pierce D (2015) No, apple isn't cutting pebble off from iOS. WIRED. http://www.wired.com/2015/04/apple-pebble-rejection/

Pogue D (2009) Some E-books are more equal than others. The New York Times, Pogue's Posts Blog. http://pogue.blogs.nytimes.com/2009/07/17/some-e-books-are-more-equal-than-others/

Savage C, Weisman J (2015) N.S.A. collection of bulk call data is ruled illegal. The New York Times. http://www.nytimes.com/2015/05/08/us/nsa-phone-records-collection-ruled-illegal-by-appeals-court.html

Schneier B (2015a) Data and goliath: the hidden battles to capture your data and control your world. Norton and Company, New York

Schneier B (2015b) Why mass surveillance can't, won't, and never has stopped a terrorist. Digg. http://digg.com/2015/why-mass-surveillance-cant-wont-and-never-has-stopped-a-terrorist
Schroeder T (2015) Freie software gegen unsere freiheit? [Free software against our freedom?]—Fighting code-abuse by military and intelligence. https://events.ccc.de/camp/2015/Fahrplan/events/7046.html
Sen AK (1999) Development as freedom. Oxford University Press
Solon O (2011) German state admits spying on citizens with trojan software (Wired UK). Wired UK
Stallman R (1983) New Unix implementation. https://www.gnu.org/gnu/initial-announcement.html
Stallman R (1999) The GNU operating system and the free software movement. In: DiBona C, Ockman S, Stone M (eds) Open sources: voices from the open source revolution, 1st edn. O'Reilly Media, Beijing
Stallman R (2013) Why Free software is more important now than ever before. https://www.gnu.org/philosophy/free-software-even-more-important.html
Stallman R (2015a) Measures Governments can use to promote free software. https://www.gnu.org/philosophy/government-free-software.html
Stallman R (2015b) Why open source misses the point of free software. https://www.gnu.org/philosophy/open-source-misses-the-point.html
Sullivan J (2014) Free software foundation statement on heartbleed vulnerability. Technical report, Free software foundation. https://www.fsf.org/news/free-software-foundation-statement-on-heartbleed-vulnerability
Wasserman H (2016) Could the 2016 election be stolen with help from electronic voting machines? http://www.democracynow.org/2016/2/23/could_the_2016_election_be_stolen
Wingfield N (2012) Apple rejects app tracking drone strikes. The New York Times Bits Blog. http://bits.blogs.nytimes.com/2012/08/30/apple-rejects-app-tracking-drone-strikes/
Zetter K (2014) It's insanely easy to hack hospital equipment. WIRED. http://www.wired.com/2014/04/hospital-equipment-vulnerable/
Zittrain J (2009) The future of the internet-and how to stop it. Yale University Press, New Haven

Chapter 6
An Intervening Ethical Governor for a Robot Mediator in Patient-Caregiver Relationships

Jaeeun Shim and Ronald C. Arkin

Abstract Patients with Parkinson's disease (PD) experience challenges when interacting with caregivers due to their declining control over their musculature. To remedy those challenges, a robot mediator can be used to assist in the relationship between PD patients and their caregivers. In this context, a variety of ethical issues can arise. To overcome one issue in particular, providing therapeutic robots with a robot architecture that can ensure patients' and caregivers' dignity is of potential value. In this paper, we describe an intervening ethical governor for a robot that enables it to ethically intervene, both to maintain effective patient-caregiver relationships and prevent the loss of dignity.

Keywords Robot mediator · Parkinson disease patients · Patient-caregiver relationship · Intervening ethical governor architecture · Rules · Evaluating the intervening ethical governor

6.1 Motivation

Robotics is currently revolutionizing various fields in our society. One particular field where the use of robotic technology is growing fast is the healthcare industry. A wide range of robot applications is being developed and successfully used in healthcare contexts such as drug manufacturing Sarantopoulos et al. (1995), robot assistants in

J. Shim (✉)
School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia
e-mail: jaeeun.shim@gatech.edu

R.C. Arkin
School of Interactive Computing, Georgia Institute of Technology, Atlanta, Georgia
e-mail: arkin@gatech.edu

M.I. Aldinhas Ferreira et al. (eds.), *A World with Robots*,
Intelligent Systems, Control and Automation: Science and Engineering 84,
DOI 10.1007/978-3-319-46667-5_6

hospitals Bohuslav et al. (1994), John (1994), and robotic surgery Howard (1992), Hockstein (2005). These applications have proven that the use of robots can improve the quality and the affordability of patient care Broadbent et al. (2009), Giulianotti (2003).

Similarly, the use of a robot can improve the quality of patient care in early stage Parkinson's disease (PD), a chronic and progressive movement disorder where symptoms continue to worsen over time. Around seven to 10 million people are diagnosed with PD worldwide, and as many as one million Americans live with PD Parkinson's Disease Foundation (2015), Willis (2010).

Over the past few years, robotic technologies have been introduced to help PD patients, mostly focused on physical rehabilitation benefits Aisen (1997), Picelli (2012). That research demonstrates that robotic training can provide benefits for preventing or delaying PD patients' loss of motor control throughout the body. Different from previous work in this domain, our research focuses on the robot's role to improve the relationship between the PD patients and their caregivers by preventing a loss of dignity (stigmatization) in PD patients and caregiver relationships Tickle-Degnen et al. (2011). One important challenge that patients generally face is the loss of control of their facial musculature, whereby patients can no longer precisely express their emotions or nuances in their face, which can leave them with blank expressions (facial masking) Müller and Stelmach (1992), Tickle-Degnen and Lyons (2004).

Since facial expression is an important social cue in human to human communication, caregivers experience difficulties in understanding the affective state of people with PD. Patients with PD are challenged in communicating with others, as facial masking prevents accurate conveyance of their emotions or feelings. Finally, facial masking worsens the quality of person-to-person interaction, giving rise to stigmatization between a caregiver and a patient, resulting in a concomitant decrease in the quality of patient care Tickle-Degnen et al. (2011). We postulate that a companion robot can remedy this challenge and reduce the communication gap between the patient and the caregiver. We aim to develop a robot mediator that can help smooth and increase the effectiveness of the interactions among PD patients and caregivers.

Stigmatization is highly related to the ethical issue of neglecting to ensure human dignity. Dignity maintenance is a chief factor in our consideration of developing a robot mediator. To reiterate, since people with PD cannot readily communicate their internal and external states due to their limited motor control, these individuals may experience the loss of dignity during therapy with their caregivers. In response, the primary goal of our robot mediator is to ensure patients' and caregivers' dignity during their interactions. To this end, robot mediators are required to intervene in patient-caregiver relationships when anyone's dignity becomes threatened.

To achieve this goal, we have developed a robot architecture that enables a robot to determine how and when to intervene when unacceptable human-human boundaries are crossed. In other work, we are using nonverbal communication to assist in maintaining an effective patient-caregiver relationship and to prevent those boundaries from being crossed in the first place Arkin and Pettinati (2014).

In this paper, we describe a robot architecture involving an intervening ethical governor that can help prevent the loss of dignity in patient-caregiver relationships. As part of developing this architecture, we define several rules for robot intervention based on evidence drawn from the medical literature and suggest ways for practically using and evaluating the model in clinical contexts. The main contributions of this paper are that it:

- Develops an ethical governor that can generate intervening actions to prevent patients' and caregivers' loss of dignity during their interactions;
- Defines necessary intervening rules based on medical literature and expert reviews; and
- Provides a novel method using focus groups for evaluating the intervening ethical governor.

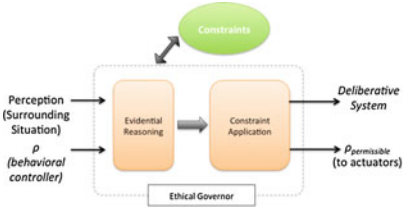
6.2 Intervening Ethical Governor Architecture

The intervening governor architecture is based on our earlier ethical governor Arkin et al. (2009, 2012), which was developed to restrict lethal action of an autonomous robot in a military context with the goal of reducing noncombatant harm Arkin et al. (2009). This original ethical governor enables a robot to evaluate the ethical appropriateness of any lethal action based on constraints derived from International Humanitarian Law.

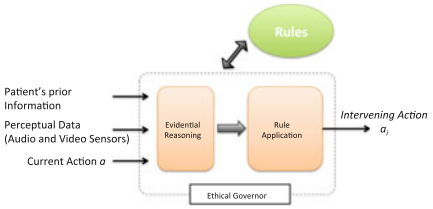
In the case of PD, the robot mediator requires a capability to determine whether it should allow continuance of the current human-human interaction (through inaction) or instead intervene by injecting itself into the relationship. In the latter case, the rules governing said intervention are derived from clinical experts or the literature Center for Substance Abuse Treatment (2006), American Psychological Association (APA) (2015), Healthcare Providers Service Organization (HPSO) (2015). Those rules will be used to determine if, when and how the robot should react to a dignity violation.

The intervening ethical governor module for the robot mediator (Fig. 6.1b) is similar to the previous ethical governor model used for controlling lethal action (Fig. 6.1a). The new model's main two components are evidential reasoning and rule application. In the evidential reasoning part, the sensory system provides data from the patients, caregiver, and the environment. Sensory data are collected and transformed into meaningful information (logical assertions) that are required for the intervention determination process. After the data is encoded, it is shared with the rule application module and generates intervening actions according to the violated rules if necessary.

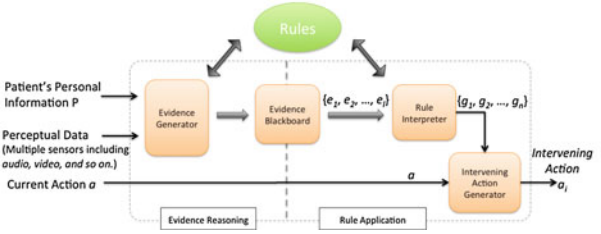
As illustrated in Fig. 6.1, we substitute the constraints in the original ethical governor with if-then rules in our intervening ethical governor model. In the original ethical governor only one overt response $\rho_{permissible}$ was possible, setting the permission-to-fire variable to True, where this response is determined by solving the constraint satisfaction problem.



(a) Original ethical governor model Arkin et al (2009)



(b) Intervening ethical governor model



(c) Detailed architecture of intervening ethical governor

Fig. 6.1 Ethical governor architectures

Different from this constraint-based model, the new intervening ethical governor requires the generation of different types of intervening actions when certain situations are violated. Therefore, the result of the intervening ethical governor will be an action, which is derived by the intervention action generator. Due to the need of multiple possible courses of action, intervening ethical governor model uses rules instead of constraints (Fig. 6.1b). Details regarding the data structures and use of rules are explained in the following section.

Figure 6.1c illustrates a more detailed view of the intervening ethical governor. Briefly, perceptual data and previous case knowledge about the patient and caregiver enter the evidence reasoning model and are encoded as evidence such as $\{e_1, e_2, \ldots, e_l\}$. Evidence is stored in the evidence blackboard (memory) that shares the information with the rule application module. The rule application module includes two components, which are the rule interpreter and the intervening action generator. In the rule interpreter, the rules are retrieved and the antecedents are mapped to the evidence values. Based on the results of comparison, if any rules are violated (i.e., they fire), the corresponding response outputs (consequents) $\{g_1, g_2, \ldots, g_n\}$ are generated for possible execution. Finally, from the set of flagged response outputs, the necessary intervening action(s) ai is generated. More detailed explanations of each component follow.

Table 6.1 Data structure for the rule

Field	Description
Type	Type of rule (obligation or prohibition)
Origin	The reference source of the rule
Activity	Indicates if the rule is currently active
Brief description	Short, concise description of the rule
Full description	Detailed text describing the rule
Logical form	Formal logical expression defining the rule
Response output	Trigger activating the intervening action when the rule is violated (fires)

6.3 Rules

Rules are the data structures that encode the intervention procedures (from experts or literature) that a robot should use to determine the correct intervening behaviors in a given situation. The data structure of rules is modified from the previous constraint data structure in the earlier ethical governor Arkin et al. (2009, 2012). Table 6.1 shows the data structure for the rule. Different from the previous constraint structure, we have one more field, which is the response output mapping to the intervening action generation mechanism.

To define intervening rules for a robot mediator, we reviewed several clinical manuals regarding how intervention should occur in patient-caregiver interaction Center for Substance Abuse Treatment (2006), American Psychological Association (APA) (2015), Healthcare Providers Service Organization (HPSO) (2015), Tickle-Degnen (2006), Tickle-Degnen (2014). From the literature, we initially generated four prohibitions and two obligations that the relationship should meet. Based on those six rules, we can provide a set of interventions for a robot mediator. Potentially, there exist more situations when intervention is required and those rules are extensible. However we currently utilize those six types since they broadly cover a range of possible cases and also can be systematically detected by a robot. More sensitive signals such as assessing change in the slight nuance of sentences are hard to detect from an autonomous agent currently, but we can add those more sensitive signals to our architecture later by developing the technology. The current pre-defined intervening rules are shown in Table 6.2 and detailed explanations of each rule are presented in the following subsections.

6.3.1 Prohibitions

We first defined three anger-related prohibitions. According to outpatient treatment (OT) rapport Center for Substance Abuse Treatment (2006), emotional excess is one

Table 6.2 Pre-defined intervening rules

<Rule> $r_{proh\_yelling}$ <type> prohibition </type> <origin> APA American Psychological Association (APA) (2015), HPSO Healthcare Providers Service Organization (HPSO) (2015), OT manual Center for Substance Abuse Treatment (2006)</origin> <active> TRUE </active> <brief description> The patient is yelling now. </brief description> <full description> Yelling shows the patient's angry emotion and it is required to be controlled by intervening in the situation. </full description> <logical form> *PatientVoiceOverThreshold* **AND** *YellingOverTimeThreshold* </logical form> <response output> g_1 </response output> </Rule>
<Rule> $r_{proh\_foulwords}$ <type> prohibition </type> <origin> APA American Psychological Association (APA) (2015), HPSO Healthcare Providers Service Organization (HPSO) (2015), OT manual Center for Substance Abuse Treatment (2006)</origin> <active> TRUE </active> <brief description> The patient is saying inappropriate words. </brief description> <full description> Foul words or insulting language are significant signals of the patient's angry emotion and intervention is necessary. </full description> <logical form> *SentenceHasFoulWords* **AND** *#ofFoulWordsOverThreshold* </logical form> <response output> g_2 </response output> </Rule>
<Rule> $r_{proh\_interrupting}$ <type> prohibition </type> <origin> APA American Psychological Association (APA) (2015), HPSO Healthcare Providers Service Organization (HPSO) (2015), OT manual Center for Substance Abuse Treatment (2006)</origin> <active> TRUE </active> <brief description> The patient is interrupting the communication. </brief description> <full description> If the patient interrupts the caregiver's communication excessively, it can be interpreted as the patient's excess of emotion. </full description> <logical form> *PatientSpeechOverlappedCaregiverSpeec* **AND** *PatientSpeechNotInBackchannel* </logical form> <response output> g_3 </response output> </Rule>
<Rule> r_{proh_quiet} <type> prohibition </type> <origin> High therapeutic rapport Tickle-Degnen (2006, 2014)</origin> <active> TRUE </active> <brief description> The patient is too quiet and he/she might be withdrawn. </brief description> <full description> If the patient is too quiet, it is difficult to establish a good communication bond between the patient and caregiver. </full description> <logical form> *PatientVoiceUnderThreshold* **AND** *QuietTimeOverThreshold* </logical form> <response output> g_4 </response output> </Rule>
<Rule> r_{oblig_stay} <type> obligation </type> <origin> High therapeutic rapport Tickle-Degnen (2006, 2014) </origin> <active> TRUE </active> <brief description> The patient should not leave their seat prior to the end of the session. </brief

(continued)

Table 6.2 (continued)

description> <full description> It is the patient's obligation to stay in therapy until the end of the session. Therefore, if the patient tries to leave the room prematurely it should be detected and an intervention generated.</full description> <logical form> *PatientUndetectedInSeat* **AND** *TimeToAbsentOverThreshold* </logical form> <response output> g_5 </response output> </Rule>
<Rule> r_{oblig_safety} <type> obligation </type> <origin> OT manual Center for Substance Abuse Treatment (2006), High therapeutic rapport Tickle-Degnen (2006, 2014)</origin> <active> TRUE </active> <brief description> Safety of the patient should be always maintained. </brief description> <full description> It is an obligation to maintain the safety in therapy until the end of the session. Therefore, any situation that can cause risk should be detected and an intervention generated. </full description> <logical form> *PatientInPotentialRisk* **AND** *CaregiverInPotentialRisk* </logical form> <response output> g_6 </response output> </Rule>

important abnormal signal from the patient, and when occurring should be intervened to re-establish positive therapeutic interactions. Especially, if patients show aggressive and angry behaviors, caregivers should intervene and try to help patients to overcome their difficulties. According to the guideline Center for Substance Abuse Treatment (2006), there are three problematic behaviors that are indicative of patients' emotional excess, which are yelling, foul language, and interrupting others.

Rule 1. Yelling: If the auditory volume of the patient is consistently over a certain threshold, a robot can determine if the patient is yelling Center for Substance Abuse Treatment (2006). For this purpose, the average decibel (dB) of the patient's voice should be measured in the first few minutes of the session. After the average decibel α is determined, the ***PatientVoiceOverThreshold*** boolean variable can be set to True when the patient's voice level is over the threshold $\alpha + \tau_{voiceDB}$ lasting a certain amount of time $\tau_{yellingTime}$ (***YellingOverTimeThreshold***). The thresholds $\tau_{voiceDB}$ and $\tau_{yellingTime}$ will be empirically set. Finally, if it is determined to be yelling, response output g_1 is transferred to the action generation component.

Rule 2. Foul language: Foul language is a significant signal showing patients' abnormal and angry emotion Center for Substance Abuse Treatment (2006). Using the speech recognition system and offensive language detection process Chen (2012), Razavi (2010), our system should determine whether the recognized sentences include foul words. Therefore, if foul words are detected (SentenceHasFoulWords) and the number of foul utterances is over the threshold τ_{foul} (***#ofFoulWordsOver Threshold***), g_2 is generated.

Rule 3. Interrupting: Interruptions from the patients can be determined by different cues including speech, hand gesture, eye gaze, and so on Lee et al. (2008).

Among those cues, speech can be used as a primary cue to determine interruptions since it is one of the most reliable cues for conflict detection. Interruption can be defined as the second speaker's unexpected speech that happens before the primary speaker's turn ends in dyadic spoken interactions Wrede et al. (2003), Grèzes et al (2013). According to this perspective, if the patient's speech overlaps to the caregiver's sentence boundaries, interruptions should be detected (***PatientSpeechOverlappedCaregiverSpeech***). In addition, even though the overlap is detected, it cannot be interruptions if the patient's speech involves backchannel utterances (uh-huh, I see, etc.). Therefore, overlapped sentences should be also evaluated whether it is backchannel (***PatientSpeechNotInBackchannel***) and if not, it can be confirmed as interruption and g_3 is generated.

Rule 4. Quiet/Withdrawn: Another prohibition rule is the withdrawn rule, which is intervention for a quiet or withdrawn patient. When a patient feels uncomfortable in joining the conversation, generally they won't speak, and caregivers recognize it as a patient's difficulty. Patients' avoidance of expression is observed especially when the therapy begins. During therapy, the caregiver's general strategies are organized around 3-components of rapport behavior Tickle-Degnen (2006), Tickle-Degnen (2014): (1) establishing mutual attentiveness and readiness to engage interpersonally, (2) establishing a positive bond between interacting parties through verbal and nonverbal positive regard/friendliness and an explicit eagerness to resolve interpersonal misunderstandings or negative interaction, and (3) flexible routines of interpersonal coordination. Because engaging a patient's attentiveness and establishing a positive bond are essential strategies when therapy starts, the lack of those components can lead to difficulty in interaction. As a result, if a patient cannot establish a positive bond with the caregiver and does not engage, it indicates a reluctance to participate.

To avoid this problem, a patient's reluctance to participate should be carefully observed. If he/she is quiet and withdrawn, it can be a signal that they don't want to continue to participate in the communication. A robot should perceive this situation and intervene by assisting in engaging the patient. For this purpose, a robot may be able to act as an "ice breaker" and help people with PD to interact with caregivers more comfortably.

The robot can recognize patients' loss of interest from different cues, where quiet is one significant signal representing patients' refusal to interact with the caregiver. If the patient's audio input is missing (***PatientVoiceUnderThreshold***) for longer than a specified threshold $\tau_{quietTime}$ (***QuietTimeOverThreshold***), the robot can flag this difficulty, and signal g_4 is transferred to the next. Sometimes, a patient's posture and eye gaze can also express their loss of interest. We can also use vision data to extract this secondary information to confirm the patient's withdrawn status.

6.3.2 Obligations

Rule 5. Stay obligation: We define the physical obligation rule for patients. When patients feel a huge challenge during therapy and try to leave the therapy room, it should be classified as a patient's attempted avoidance of the difficult situation. It is another important moment when a robot should intervene and help patients to re-engage in the relationship with the caregiver. If the patient leaves the sitting position, it should be detected by the robot. A robot can observe the patient's position via different sensors such as a camera or a pressure sensor in the seat. In our system, the seat sensor will be placed in the patient's seating location. The seat sensor determines if a person occupies a seat by detecting pressure, and therefore the system can recognize whether the patient leaves using this sensor (***PatientUndetectedInSeat***). However, the seat sensor can incorrectly determine that the patient is absent even though they do not intentionally try to leave the position. For example, if patients try to reposition their posture, their pressure might be under-detected. To avoid these problems, the system will determine whether the absence is maintained over a certain time threshold $\tau_{absencetime}$ (***TimeToAbsentOverThreshold***). If it is over this condition, it will be determined as a violation and the signal g_5 is generated.

Rule 6. Safety-first obligation: Safety is always the most important factor in any clinical situation. As such, during therapy sessions, if any situations that violate the safety of patients are detected, then an appropriate intervening action should be generated. As shown in rule r_{safety}, if the situation is determined as one that could pose risk (***PatientInPotentialRisk*** OR ***CaregiverInPotentialRisk***), signal g_6 is transferred to the r_{oblig_safety} action generator in order to bring about the intervening action. The violation of a safety situation can be determined by the pre-encoded set of risk situations. As shown in Fig. 6.2, the patient's prior/personal information is an

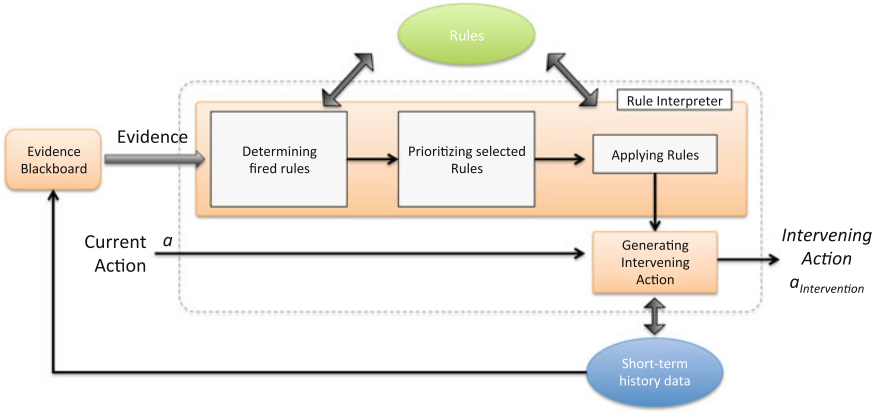


Fig. 6.2 Detailed architecture of the rule application

input of our intervening ethical governor system. Generally, it includes diagnosis of the patient's medical history that can evidence for potential risks. For example, the diagnosis may contain patients' prior emergency experiences, so it can be encoded as a risk in the system. A doctor's recommendations or instructions to avoid any potential risks are also generally stated. Therefore, when the system is initialized, this prior information will be reviewed and encoded to the set of risks in the system specific to the current patient. It should be noted that the privacy of the patient information will be guaranteed by storing and managing in a secured way. Finally, by comparing the current perceived situation to those encoded risks, the violation of a safety can be determined.

6.4 Evidential Reasoning and Rule Application

6.4.1 Evidential Reasoning

The evidential reasoning process transforms incoming perceptual and prior data into evidence in the form of logical assertions to be used by the rule application process. In this process, audio, video, and other sensory data from the patient, caregiver, and environment are perceived, interpreted, and transferred to the evidence generator. Audio data from the patient and the caregiver will be collected through microphones. Pressure data will be also gathered from the patient's seat sensor. This sensor data is then converted into situation-specific percepts.

After extracting the perceptual data from the sensory raw data, it is used by the evidence generator to create logical assertions describing the current state of the patients and caregivers. The necessary Boolean logical assertions, which are used as the evidence $e \in E$, are defined by the active rules. From the intervention rules described in Sect. 6.3, we can determine the current set of evidence E as follows:

Set of evidence E = {PatientVoiceOverThreshold, YellingOverTimeThreshold, SentenceHasFoulWords, #ofFoulWordsOverThreshold, PatientSpeechOverlappedCaregiverSpeech, PatientSpeechNotInBackchannel, PatientVoiceUnderThreshold, QuietTimeOverThreshold, PatientUndetectedInSeat, TimeToAbsentOverThreshold, PatientInPotentialRisk, CaregiverInPotentialRisk}.

In the evidence generator, evidence is calculated by each relevant algorithm using the appropriate perceptual data, and prior information. The evidence is then used to determine if any of the active rules apply. The evidence is stored and updated in the evidence blackboard, which serves as the communication medium between the evidential reasoning process and the rule application process.

6.4.2 Rule Application

The rule application process (Fig. 6.2) is responsible for reasoning about the active rules and evaluating if any intervening behavior by the robot is required. Figure 6.2 illustrates the steps involved in applying the rules. By observing evidence from the evidence blackboard, the process first determines which rules are currently being fired. As we explained above, if the antecedent of any specific rule is calculated as *TRUE* based on the current evidence, it is considered as an interventional rule and its response output g_i is generated. If several rules are determined to be active, these rules are prioritized in an order predetermined by an expert's input. For example, we received an expert comment stating, "Safety should always be first among all the rules," so it is assigned the highest priority. More than one action may be generated if there is no conflict on the robots actuators. If two or more rules of equal priority apply, one will be randomly chosen. By applying the selected rules according to priority, an intervening action(s) is generated in the intervening action generator.

The prioritized list of response outputs is used as input for the intervening action generator. According to the order of response outputs in this list, the associated action set is fetched to determine the intervening action(s). Currently four action sets are defined: a_{angry}, a_{quiet}, a_{stay}, and a_{safety}. Action set a_{angry} is associated with response outputs g_1, g_2, and g_3, and a_{quiet} is with g_4. Response outputs of the obligation rules g_5 and g_6 are correlated to a_{stay} and a_{safety}. For example, if the prioritized list $\{g_6, g_2\}$ is an input of the intervening action generator, action set a_{safety} is first fetched to generate the intervening action, followed by action set a_{angry}.

Each action set, described in Table 6.3, contains one or more potential verbal and nonverbal cues based on medical literatures Center for Substance Abuse Treatment (2006), American Psychological Association (APA) (2015), Healthcare Providers Service Organization (HPSO) (2015). To generate the final intervening action, typically one specific verbal and nonverbal cue is selected and combined. When more than one action cue is possible, the intervening action generator will randomly select one verbal cue and one nonverbal cue and combine those two cues together to be performed as the intervening action. Next any remaining action sets are reviewed, and if no conflict exists, they are combined into the final intervening action Arkin et al. (2003). For a robot mediator, only one verbal cue is performed at a time and selected from the highest prioritized action set. However, several nonverbal cues can be integrated into one intervening action if there is no conflict. The system evaluates if any nonverbal cues from the next priority action set in order that do not conflict with the current action's actuators, and if they exist, these nonverbal cues are integrated into the final action.

Intervening action set a_{angry}: Action set a_{angry} contains intervening verbal and nonverbal cues to handle angry patients. According to the OT rapport Center for Substance Abuse Treatment (2006), angry patients should be treated as follows: "Identify specific behaviors that are inappropriate. State that these behaviors are not allowed. Identify the consequences if the behaviors continue. (Center for Substance Abuse Treatment (2006), p. 147)" In addition, other clinical manuals American Psycholog-

Table 6.3 Verbal and nonverbal cues for each action

a_{angry} { <Verbal cues> <V1> "Oh, are you upset little bit?" <V2> "Let's calm down now." <Nonverbal cues> <N1> Turn the head to the patient. <N2> Other appropriate (down) hand gesture. }
a_{stay} { <Verbal cues> <V1> "The session is not yet finished!" <V2> "Please follow me! Let's go back together." <Nonverbal cues> <N1> Turn the head to the patient. <N2> A robot points to the therapy place. <N3> Other appropriate hand gesture. }
a_{quiet} { <Verbal cues> <V1> "You are not speaking a lot today?" <V2> "Could you please answer those questions?" <V3> Appropriate Jokes <Nonverbal cues> <N1> Turn the head to the patient. <N2> Other appropriate (cheering) hand gesture. }
a_{safety} { <Verbal cues> <V1> "I think it's not safe. Let's stop it now." <Nonverbal cues> <N1> Turn the head to the patient. <N2> Other appropriate (stopping) hand gesture. }

ical Association (APA) (2015), Healthcare Providers Service Organization (HPSO) (2015) state strategies for how to handle the anger; (1) keep looking for anger signs, (2) show empathy, and (3) remain calm and professional. Based on those manuals, the intervening action cues can be defined as a_{angry} in Table 6.3.

Intervening action set a_{quiet}: For quiet and withdrawn patient, a robot mediator should help him/her join the conversation with more relaxed feelings. To intervene the patient trying to avoid participating in the conversation, action set a_{quiet} in Table 6.3 should be used. Especially, this intervening action can be helpful as an icebreaker when the relationship begins.

Intervening action sets a_{stay} **and** a_{safety}: Response output g_5 indicates that a patient is currently trying to leave (stay-obligation). When this situation occurs, a robot mediator should warn and try to re-engage the patient (a_{stay} in Table 6.3). When safety in a situation is violated, response output g_6 is triggered. To negate those possible risks, robot mediators should warn patients and caregivers and immediately request to stop the current process (a_{safety} in Table 6.3).

6.4.3 Short-Term History Data

If a specific intervening action is generated from the action generator yet the same violation is repeatedly detected by a robot mediator, then this action is deemed ineffective for the current case and should therefore not be repeated. Previous intervening actions, latency, and performance are maintained as ***Short-term history data*** and this data is used to filter out ineffective actions in the intervening action generator.

6.5 Evaluating the Intervening Ethical Governor

The main contribution of this paper is the presentation of a novel ethical governor that can determine and generate appropriate intervening actions for a robot mediator in the patient-caregiver relationship. As a first step in evaluating our governor, a PD expert in occupational therapy reviewed our predefined intervention rules, the results of which guided the modification of intervention rules and actions. Some highlights of the review include:

- The safety-first obligation should take priority over all other rules.
- Intervening actions need to be modified so that they do not blame patients. Although they are generated based on OT manuals and other medical literatures, those instructions can be sensitive and need to be regulated for PD patients.
- A patient's prior diagnosis or personal information can be more important in PD cases and should be integrated into the rules.

The current model of the intervening ethical governor and intervention rules resulted from modifications made according to those comments. Next, the intervening ethical governor is being applied to a robot mediator and will then evaluated by focus groups of PD patients and caregivers. A specific task (e.g., weekly medication-sorting) will be selected for patient-caregiver interaction and a robot mediator placed during the task to assess and if necessary perform intervening actions. We will generate several stress-generating scenarios that can prompt different intervening actions and record them as simulation videos to be reviewed by focus groups. In addition, by evaluating the intervening ethical governor, we expect to add and/or delete intervention rules and modify current rules based on expert knowledge. Finally, we anticipate evaluating the system in an actual clinical setting.

6.6 Conclusions

We introduce an intervening ethical governor that enables a robot mediator to generate appropriate intervening actions in PD patient-caregiver interactions. Using these intervening actions, we aim to produce a robot mediator that can improve PD patient-caregiver communication and relationships. In this context, the overarching goal of

the governor is to maintain dignity in human interactions by using robotic technology. In the model of the intervening ethical governor, six intervening rules are defined based on medical literature. To validate the system, those rules are reviewed by PD experts and modified. We next apply the governor to our robot mediator and simulate a robot in PD patient-caregiver interactions with a specific task. Several situations in which robots can generate intervening actions are simulated and recorded, the videos of which are reviewed by focus groups and evaluated to inform the modification of intervention rules.

Acknowledgements This work is supported by the National Science Foundation under Grant #IIS 1317214 in collaboration with Profs. Linda Tickle-Degnen and Matthias Scheutz at Tufts University.

References

Aisen M (1997) The effect of robot-assisted therapy and rehabilitative training on motor recovery following stroke. Arch Neurol 54(4):443–446

American Psychological Association (APA) (2015) Controlling anger before it controls you. http://www.apa.org/topics/anger/control.aspx

Arkin R, Pettinati M (2014) Moral emotions, robots, and their role in managing stigma in early stage Parkinson's disease caregiving. In: Proceedings of workshop on new frontiers of service robotics for the elderly, RO-MAN

Arkin R, Fujita M, Takagi T, Hasegawa R (2003) An ethological and emotional basis for human-robot interaction. Robot Auton Syst 42:3–4

Arkin R, Ulam P, Duncan B (2009) An ethical governor for constraining lethal action in an autonomous system. Technical report, Georgia Tech, Technical Report (No. GIT-GVU-09-02)

Arkin R, Ulam P, Wagner A (2012) Moral decision-making in autonomous systems: enforcement, moral emotions, dignity, trust and deception. Proc IEEE 100(3):571–589

Bohuslav Z, Voss H, Fincati A (1994) Robotic drug dispensing system. U.S. Patent No. 5, 341,854

Broadbent E, Stafford R, MacDonald B (2009) Acceptance of healthcare robots for the older population: review and future directions. Int J Soc Robot 1(4):319–330

Center for Substance Abuse Treatment (2006) Substance abuse: clinical issues in intensive outpatient treatment

Chen Y (2012) Detecting offensive language in social media to protect adolescent online safety. In: Procs. of IEEE International conference on social computing, pp 71–80

Giulianotti P (2003) Robotics in general surgery: personal experience in a large community hospital. Arch Surg 138(7):777–784

Grèzes F, Richards J, Rosenberg A (2013) Let me finish: automatic conflict detection using speaker overlap. In: Proc Interspeech, pp 200–204

Healthcare providers service organization (HPSO) (2015) Handling the angry patient. http://www.hpso.com/resources/article/2.jsp

Hockstein N (2005) Robotic microlaryngeal surgery: a technical feasibility study using the daVinci surgical robot and an airway mannequin. Laryngosc 115(5):780–785

Howard P (1992) Development of a surgical robot for cementless total hip arthroplasty. Clin Orthop Relat Res 285:57–66

John E (1994) HelpMate: an autonomous mobile robot courier for hospitals. In: Proceedings of IROS'94

Lee C, Lee S, Narayanan S (2008) An analysis of multimodal cues of interruption in dyadic spoken interactions. In: Proceedings of interspeech, pp 1678–1681

Müller F, Stelmach G (1992) Pretension movements in Parkinson's disease. Adv Psychol 87:307–319

Parkinson's Disease Foundation (2015) http://www.pdf.org/en/parkinson_statistics

Picelli A (2012) Robot-assisted gait training in patients with Parkinson disease a randomized controlled trial. Neurorehabilitation Neural Repair 26(4):353–361

Razavi A (2010) Offensive language detection using multi-level classification. In: Advances in artificial intelligence. Springer, pp 16–27

Sarantopoulos P, Tayfur A, Elsayed A (1995) Manufacturing in the pharmaceutical industry. J Manuf Syst 14(6):452–467

Tickle-Degnen L (2006) Nonverbal behavior and its functions in the ecosystem of rapport. In: SAGE handbook of nonverbal communication, SAGE, pp 381–399

Tickle-Degnen L (2014) Therapeutic rapport. In: Radomski M, Trombly Latham C (eds) Occupational therapy for physical dysfunction, 7th edn. Wolters Kluwer, pp 412–427

Tickle-Degnen L, Lyons K (2004) Practitioners's impressions of patients with Parkinson's disease: the social ecology of the expressive mask. Soc Sci Med 58(3):603–614

Tickle-Degnen L, Zebrowitz L, Ma H (2011) Culture, gender and health care stigma: practitioners' response to facial masking experienced by people with Parkinson's disease. Soc Sci Med 73(1):95–102

Willis A (2010) Geographic and ethnic variation in Parkinson disease: a population-based study of US Medicare beneficiaries. Neuroepidemiology 34(3)

Wrede B, Shriberg E, Spotting (2003) "hot spots" in meetings: human judgments and prosodic cues. In: Proceedings of Eurospeech, pp 2805–2808

Chapter 7
Exploring the Ethical Landscape of Robot-Assisted Search and Rescue

Maaike Harbers, Joachim de Greeff, Ivana Kruijff-Korbayová, Mark A. Neerincx and Koen V. Hindriks

Abstract As robots are increasingly used in Search and Rescue (SAR) missions, it becomes highly relevant to study how SAR robots can be developed and deployed in a responsible way. In contrast to some other robot application domains, e.g. military and healthcare, the ethics of robot-assisted SAR are relatively under examined. This paper aims to fill this gap by assessing and analyzing important values and value tensions of stakeholders of SAR robots. The paper describes the outcomes of several Value Assessment workshops that were conducted with rescue workers, in the context of a European research project on robot-assisted SAR (the TRADR project). The workshop outcomes are analyzed and key ethical concerns and dilemmas are identified and discussed. Several recommendations for future ethics research leading to responsible development and deployment of SAR robots are provided.

Keywords Roboethics · Search and rescue robotics · Human-robot interaction · Ethical concerns · Values · Value sensitive design

M. Harbers (✉) · J. de Greeff · K.V. Hindriks
Delft University of Technology, Delft, The Netherlands
e-mail: M.Harbers@tudelft.nl

J. de Greeff
e-mail: J.Degreeff@tudelft.nl

K.V. Hindriks
e-mail: K.V.hindriks@tudelft.nl

I. Kruijff-Korbayová
Language Technology Lab, DFKI, Saarbruecken, Germany
e-mail: ivana.kruijff@dfki.de

M.A. Neerincx
TNO Human Factors, Delft University of Technology, Delft, The Netherlands
e-mail: mark.neerincx@tno.nl

M.I. Aldinhas Ferreira et al. (eds.), *A World with Robots*,
Intelligent Systems, Control and Automation: Science and Engineering 84,
DOI 10.1007/978-3-319-46667-5_7

7.1 Introduction

With advancements in AI and robotics, robots that share an environment and interact with people are becoming ubiquitous. This development has fueled a growing realization that ethics of human-robot interaction needs to be addressed, evidenced by a growing number of publications, workshops and conferences addressing roboethics (Wallach and Allen 2008; Lin et al. 2011; Murphy and Woods 2009; Malle et al. 2015). As Riek and Howard (2014, page 5) put it "One especially wants to avoid giving the impression that it is the responsibility of the ethicist to instruct scientists and engineers on what they may and may not do. Ethics should, instead, be understood as making a constructive contribution to work in HRI". To foster this development, a code of ethics and practical guidelines have been proposed for robot engineers and HRI practitioners (Ingram et al. 2010; Riek and Howard 2014; Murphy and Woods 2009).

For some domains, ethical concerns regarding the application of robots have received a lot of attention (Łichocki et al. 2011), e.g. in the military domain, car industry, healthcare and education. However, in the field of Search and Rescue (SAR) it appears that ethics related to the use of robots has not so much been addressed; indeed, it is telling that in the EURON Roboethics Roadmap (Veruggio 2006) SAR robots are only mentioned as a subcategory of 'outdoor robots'. Ethics and values are relevant for responsible development (e.g., requirements) and deployment (e.g., working agreements) of SAR robots. In this paper, we therefore provide an exploration of the ethical landscape surrounding robot-assisted SAR missions.

We explore the ethical robot-assisted SAR landscape by identifying and analyzing humans values (e.g. trust, autonomy and privacy) and value tensions. Value tensions refer to situations in which technology supports one value while at the same time hinders another; as such they are indicators of potential ethical dilemmas. Our approach is inspired on the Value Sensitive Design (VSD) methodology, which accounts for human values throughout the design process (Friedman et al. 2013). We conduct a series of three Value Assessment workshops with SAR workers—in this case firefighters—in which we make use of VSD methods to assess and analyze the stakeholders and their values in the SAR field. Using the workshop results, literature and experiences in the TRADR project, we identify key ethical concerns and dilemmas for the robot-assisted SAR field.

In this paper we first provide a description of the robot-assisted SAR domain. Then, we briefly summarize ethical themes in different robot application fields that are relevant to SAR. We then describe the setup and execution of the Value Assessment workshops, and the workshop outcomes (stakeholder values and value tensions). From these outcomes, we derive and discuss several main ethical concerns and dilemmas specific to robot-assisted SAR.

7.2 Background

7.2.1 Robot-Assisted Search and Rescue

The SAR domain is a unique area of application because it inherently entails an unstructured (often destructed) environment, that is commonly hazardous for both people and equipment. Particularly for first-response missions, work is typically done under time pressure (every minute can count) and in harsh conditions; this can lead to physical and/or mental strains on rescue workers, increasing the risk of developing psychiatric and post-traumatic distress (Fullerton et al. 1992; Chang et al. 2003; Bos et al. 2004). SAR can also happen over prolonged periods of time (days, weeks, months) as part of ongoing disaster response. These domain characteristics entail some specific ethical considerations, e.g. what is morally acceptable to ask from rescue workers in terms of mental and physical well-being when lives are at stake. Work specifically addressing the ethics of disaster response is hard to find, but some guidelines exists, e.g. the Council of Europe's "Ethical principles on disaster risk reduction and people's resilience" (Prieur 2012) specifically dictates how rescue workers should behave ethically, as well as specifying that rescue workers should have access to psychological assistance during and after disaster response missions. Other work has discussed the ethics of disaster management (Geale 2012), drawing parallels with ethics of humanitarian aid, while others specifically address ethics of firefighters (Sandin 2009), comparing it with the ethics of the medical profession.

Generally, the SAR domain is perceived as an application area in which robots can provide a valuable contribution. Robots are capable of traversing areas that are inaccessible for humans, may carry elaborated sensory equipment beyond human capabilities (e.g. infrared) and can provide unique perspectives (e.g. aerial view) contributing to situation awareness. There exists a large body of research addressing the employment of SAR robots (Murphy 2014), along with actual application in the field (Murphy 2004; Murphy et al. 2012).

The types of robots that are employed in SAR environments are quite various; common types include Unmanned Ground Vehicles (UGV), carrying a variety of sensors (e.g., laser range finder, video, audio, infrared) and typically equipped with tracks to navigate unstructured terrains (see Fig. 7.1), and Unmanned Aerial Vehicles (UAV) that can provide high-level (aerial) view of the disaster area. Generally, robots are employed in search areas that are inaccessible for humans because they are too dangerous, or because of physical constrains (too small, too high). As of today, robots are always controlled by human operators (human-in-the-loop), but some functionality is becoming (partially) autonomous (Birk and Carpin 2006; Okada et al. 2011; Zuzánek et al. 2014).

Interaction between humans and SAR robots can happen in a number of distinct manners. Humans interact with robots as operators, as infield-rescuers or as victims, each yielding different types of HRI. For instance, robots controlled by an operator are embedded within a clear hierarchical structure, but when robots encounter a

Fig. 7.1 The TRADR UGV operating during an exercise aimed to capture a SAR context

victim during a mission, the human is in some way dependent on the robot, e.g. the robot supports evacuation of victims by lifting and carrying them.

7.2.2 Roboethics

There are many sorts and types of robots (e.g. military, surveillance, service, educational and entertainment robots), and they are used in diverse application domains. Different robot types and application domains pose their own design challenges and ethical concerns (Łichocki et al. 2011).

There is some work addressing robot ethics against a backdrop of the SAR domain. For instance, Kruijff and Janíček (2011) propose a method of modeling accountability in human-robot teams, thus endowing artificial systems (robots) with some form of moral accountability. However, to the best of our knowledge there is relatively little work explicitly addressing the ethics of robot-assisted SAR. In order to get a better grip on the ethical concerns regarding SAR robots, we therefore discuss ethics surrounding the use of robots in the healthcare and military domain. There is a considerable amount of work on roboethics in these domains, and both have links with the SAR domain (healthcare resembles victim care in SAR, and military as well as SAR robots are used in rough and unknown terrains to perform reconnaissance and search for targets).

Robots in healthcare are used for different tasks, which are often categorized as monitoring, housekeeping, and companionship tasks (van Wynsberghe 2013; Sharkey and Sharkey 2012; Decker 2008; Butter et al. 2008). Monitoring tasks involve, for instance, keeping track of someone's physical activities or medicine intake, and detecting abnormal or dangerous situations. Housekeeping and assistance tasks include cleaning, washing, carrying objects, and serving food and drinks. Examples of companionship activities are displaying emotions, responding to emotion and touch, talking and playing. An ethical concern related to these robot activities includes the issue of responsibility for a robot's (failed) actions, e.g. who is responsible if a robot harms a patient or provides wrong medical advice? Another concern involves privacy, e.g. who has access to the data that a robot collects about a patient, and under what circumstances? Also, what consequences does a robot that serves as a companion have for human-human contact? It is particularly important to address these issues because healthcare robots often interact with vulnerable groups of people such as patients, elderly or children.

Military robots serve on the ground as stationary robots or unmanned ground vehicles (UGVs), in the air as drones, unmanned aerial vehicles (UAVs) or remotely piloted systems, and on or under water as unmanned ships or submarines. Task performed by military robots include monitoring, navigating, carrying, target tracking and firing. Proponents of military robots see them as a way to relieve humans from dull, dirty and dangerous tasks, and sometimes also as a way to improve performance (Arkin 2009). Others, however, are concerned about shifting control from humans to robots, in particular if this includes the application of lethal force (Lucas Jr 2011; Sparrow 2007). Nowadays, most military robots are still tele-operated by human operators, but the development towards more automation has been debated in media, politics and academia, most notably in the form of the campaign 'Stop Killer Robots' (Docherty 2012). These discussions concern issues of responsibility for robot behavior, and psychological effects of military robot use on the enemy, robot (drone) operators and civilians in war zones (Lin et al. 2009).

7.3 Value Assessment Workshops

We performed three workshops to assess values that play an important role in robot-assisted SAR. The workshops were conducted during an end-user meeting of the TRADR project that was held in Pisa in September 2014. The Value Assessment workshops were inspired on Value Sensitive Design (VSD) methods. In this section we provide some background on the TRADR project and VSD; subsequently, we describe the setup of the workshops and provide the results.

7.3.1 The TRADR Project

The *Long-Term Human-Robot Teaming for Robot-Assisted Disaster Response* (TRADR) project[1] aims to develop robots that are able to provide assistance during disaster response missions, working alongside human rescue workers as teammembers rather than as tools (Kruijff-Korbayová et al. 2015). A TRADR team is comprised of human rescue workers (team leader, robot operators, infield rescuer), UGVs and UAVs. Three fire-fighting brigades (Dutch, German and Italian) represent SAR end-users and are part of the TRADR consortium. A yearly development cycle—including exercises and evaluations with end-users—contributes to align the project towards employment in the field.

As a disaster response mission may last days, months or even years, within TRADR there is an emphasis on building persistent models of the environment, multi-robot action and human-robot teaming. Towards this end, both low-level robot control aspects and higher-level human-robot teaming aspects are addressed. Particularly the latter entails a (re)definement of the roles that robots may play within a search and rescue team. A robot's assigned role, its capabilities, its appearance and its behavior will influence expectations that people interacting with the robots have. As the role that robots play in SAR teams changes—e.g. by enabling robots to act responsibly in a team and by endowing them with social intelligence (Fincannon et al. 2004)—moral expectations may become heightened, potentially up to levels not achievable by the robots.

7.3.2 Value Sensitive Design

VSD is a "theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process" (Friedman et al. 2013). VSD defines values as "the things that people or groups of people consider important in life".

Important concepts in VSD are stakeholders, values and value tensions. A distinction is made between direct and indirect stakeholders of a system. Direct stakeholders interact directly with the system or its output, and indirect stakeholders are impacted by the system without interacting with it directly. Stakeholders have values. Values that play a role in the design of technology are, for instance, autonomy, security, privacy, safety, trust, responsibility, sustainability, and fun. Value tensions occur when a particular design of a system supports one value, but hinders another. For example, supporting the value of security, e.g. by placing more surveillance cameras, may hinder privacy.

VSD contains a rich collection of methods and techniques that allow designers to account for human values throughout the design process. There are methods to e.g., identify and analyze value tensions (Miller et al. 2007), promote envisioning of

[1] http://www.tradr-project.eu.

long-term influence of new technology (Friedman and Hendry 2012), and identify requirements that account for values (Harbers et al. 2015).

7.3.3 Setup of the Workshops

Three Value Assessment workshops were organized to explore values at stake in the robot-assisted search and rescue domain. The specific order of the activities in the Value Assessment workshops is new, but all activities are based on existing VSD methods.

The workshop participants were all professional firefighters from the *Corpo Nazionale Vigili del Fuoco*, the Italian national firefighting organization. They were from different levels in the organization (both officers and field workers), and none of them worked with robots in their daily work. The three workshops had 12, 8, and 3 participants, respectively.

All three workshops contained the following three steps: (a) identify stakeholders, (b) identify values for each stakeholder (group), and (c) examine the relation between stakeholder values and search and rescue robots. Step (a) involved the identification of both stakeholders that interact with the robots and those that do not interact with them, but are affected by the robots. In step (b), values were identified for each stakeholder, where values had to be important for the stakeholders in that specific role. For instance, a fire-fighter may value both safety and friendship, but in the role of a fire-fighter, safety is more important than friendship. In step (c), the relation between values and technology was considered by identifying positive and negative effects of search and rescue robots on each stakeholder value. This last step serves to provide context to the stakeholder values, and make clear how they can be affected by SAR robots.

Due to practical reasons, the three workshops differed in duration (4, 2, and 2 h, respectively), and because of that, only the participants of the first workshop identified value tensions (in addition to that, they also created a mind map of "Disaster Response" and prioritized the importance of different values). Although there were some differences between the workshops in number of participants, duration and activities, we believe this had no particular impact on the results, and therefore we aggregated the outcomes over all workshops.

7.3.4 Workshop Outcomes

Table 7.1 provides an overview of the stakeholders and values that were identified in the workshops. Other relevant values and stakeholders—that were not mentioned by the workshop participants—may exist, but here we only report the workshop outcomes. Regarding the relation between technology and stakeholder values, considerably more positive than negative effects of SAR robots on stakeholder values

Table 7.1 Workshop outcomes

Stakeholder	Values
Firefighter	Personal safety, safety of others, access to information, well-being, effectiveness, ease of use, authority
Victim	Personal safety, health, well-being, access to information, contact
Paramedic	Personal safety, access to information, contact, health, well-being
Policemen	Personal safety, security, neutrality, effectiveness, courage, security, trust, access to information
Press	Impartiality, transparency, access to information
Local authorities	Access to information, sharing information, safety, healthy finances
Observers	Curiosity, safety
Electricity company	Access to information, safety

were identified. We provide a few representative examples of relations between stakeholder values and technology below.

Personal safety, for instance, was identified as a value of both firefighters and victims. For firefighters, only positive effects of SAR robots on this value were identified: SAR robots make it possible for firefighters to stay away from dangerous situations. For victims, however, positive and negative effects were identified. On the one hand, it was said that robots can find and rescue victims, but on the other hand, they may be dangerous for victims, e.g. if they have inflammable batteries or when they fly or drive into a human.

Another example is the value of health for paramedics. This value is supported by SAR robots in the sense that robots can provide information about the physical state of victims, e.g. blood circulation, breath, and heart rate. But at the same time, the value is hindered because robots cannot provide health information about victims of the same quality as a human would provide.

A final example is that SAR robots were also thought to have a positive effect on local authorities' value of sharing information; e.g., robots allow local authorities to provide more information about the situation to press, citizens and family of victims.

7.3.5 Value Tensions

Value tensions involve conflicts between values of different stakeholders groups, values of one stakeholder group, or one value of one stakeholder group (which can become threaded in the wake of introducing new technology) (Miller et al. 2007). In this subsection, we discuss the value tensions regarding the deployment of SAR robots that were identified by the workshop participants. As such, they represent the stakeholders' view and indicate where potential conflicts—that are important from their perspective—may arise. Some scenarios—e.g. what happens when SAR robots

are armed—are hypothetical and unlikely to occur; we nevertheless include these as they are part of the result.

Hindering versus supporting safety. Robots can both support and hinder the safety of the people that encounter them, such as victims and rescue workers. On the one hand, robots can improve the search and rescue operation. But on the other hand, they can be dangerous, for instance, when they fail to identify a human being and collide (flying or driving) with the human. Also, equipping robots with weapons and ammunition may support the safety of search and rescue workers or policemen, but may hinder the safety of victims or other people encountered by the robot.

Safety versus well-being. The deployment of robots can support safety of victims by making the SAR operation faster and more effective, but it can hinder the victims' well-being. For example, it may be a shocking experience to be trapped, wounded and lost, and suddenly be confronted with a robot, in particular, if there are no humans around. There may also be victims that do not want to be saved by a robot.

Effectiveness of firefighter versus police. SAR robots can be deployed for a lot of different activities. When there is a limited amount of robots, choices have to be made regarding their deployment. In such situations, for instance, deploying a robot for activities of the fire brigade hinders effectiveness of policemen, and vice versa. This tension may also occur within one stakeholder group, e.g. firefighters, when the group is divided into sub-teams, and there are not sufficient robots for all sub-teams.

Transparency versus privacy. Robots make it possible to collect more information of a disaster through their cameras and other sensors. Transmitting this information to the press supports transparency, as it allows the press to better inform the public about the situation at hand. However, it may happen that privacy sensitive information about victims is spread this way, e.g., when family members learn about a victim's situation through media rather than through personal conversation.

Safety and effectiveness versus healthy finances. Deployment of robots can increase the safety and effectiveness of rescue workers during a disaster response situations. However, the purchase of robots may be expensive and hinder the local authorities value of healthy finances.

Transparency and access to information versus well-being. Robots make it possible to collect more information of a disaster. Spreading this information can support transparency and access to information for the public ad other stakeholders. But at the same time, it may hinder well-being by scaring people and creating unnecessary panic.

7.4 Ethical Concerns and Dilemmas in Robot-Assisted SAR

In the previous section we described the outcomes of the Value Assessment workshops. In this section we combine the workshop outcomes with insights obtained from our work in TRADR project (e.g. from interacting with different stakeholders,

Table 7.2 List of ethical dilemmas identified for robot-assisted SAR

#	Dilemma
1	Should SAR robots be employed when they might help saving lives, but their application might also lead to casualties?
2	Should one develop SAR technology that is intended for peaceful purposes even when it has clear military potential?
3	Should one replace infield workers by robots if that leads to suboptimal performance?
4	To what extent should information collected by robots be processed to make it more digestible, at the risk of losing or misrepresenting information?
5	Should one deploy robots, knowing that this may raise false expectations and runs the risk of degraded performance?
6	Should one deploy robots that may yield responsibility assignment problems?

observing them when they interact with robot technology) and insights obtained from the literature, identifying risks for the SAR domain that can lead to ethical dilemmas. We do that by grouping similar concerns (potential negative effects of technology on values and value tensions), and then including those concerns that either turned up multiple times, e.g. within the workshops or in both a workshop and in the literature, or that are considered essential based on literature on roboethics.

Our analysis results in the following six main ethical concerns that are relevant to the SAR domain: (1) safety risks, (2) decreased performance due to replacement of humans, (3) loss of relevant information, (4) false expectations about robot capabilities, (5) loss of privacy, and (6) responsibility assignment problems. The first five concerns are directly derived from the workshop outcomes, though we rephrased some of them to make them more generally applicable to SAR robots. We added the last concern, responsibility assignment problems, as it is often discussed in literature on roboethics (e.g. (Noorman and Johnson 2014)), and it may also apply to SAR robots. All of the concerns may also apply to the use of robots in other domains, but in this section we discuss how they apply to the SAR domain specifically. In addition, we highlight ethical dilemmas related to these risks (listed in Table 7.2).

Safety risks. One of the main objectives of SAR is to bring people into safety, and the field inherently has to deal with safety risks. SAR robots can reduce a lot of these risks, most notably, when robots instead of rescue workers explore dangerous areas to search for and rescue victims. The introduction of SAR robots, however, also yields new safety risks, where we make a distinction between safety risks *within* and *beyond* a single search and rescue mission.

Risks due to robot use within a SAR mission are caused by potential malfunctioning, or otherwise inappropriate behavior of the robot. A SAR robot, for instance, can break down, cause collisions in unstable buildings, or drive or fly into a human. Even if the robot is technically performing sound, its behavior can still be harmful due to its interaction with the environment. This poses dilemma #1: *should SAR robots be*

employed when they might help saving lives, but their application might also lead to casualties?

Safety risks that reach beyond the scope of single missions are related to possible dual use of SAR robots. The technology developed for robot-assisted SAR—while not being intended to—is often also applicable in military domains, where the aim may be killing rather than rescuing people. This is the case because characteristics of SAR missions are to a large extent very similar to war zone missions, i.e. performing reconnaissance, providing tactical overview, searching for persons using a variety of sensors and information sources in rough, unstable and unpredictable terrains. Thus, this poses dilemma #2: *should one develop SAR technology that is intended for peaceful purposes even when it has clear military potential?*

Decreased performance due to replacement of humans. The application of SAR robots can lead to a reduction in the number of human (infield) rescue workers. As of to date, robots are generally perceived as an addition to SAR missions. But once a technology is in place, it is not inconceivable that robots—in certain situations—may be used as substitutes for, rather than additions to, human rescue workers. This is likely to happen especially in those situations that pose high risks on human rescue workers, but are currently deemed acceptable.

Replacement of humans by robots may lead to degraded performance with respect to victim contact, situation awareness, manipulation capabilities, etc. For instance, a robot may scare a victim who is not expecting a robot, or does not recognize it as a benign SAR robot. Even though the robot would be equipped with social capabilities or mediate contact between a victim and rescue workers at a distance, it would probably not be able to calm the victim as much as a human would.

Another example of degraded performance due to robots replacing rescue workers is that mediated contact may make it harder for medical personnel to perform triage or provide medical advice and support. The potential replacement of human workers by robots yields the following dilemma #3: *should one replace infield workers by robots if that leads to suboptimal performance?*

Loss of relevant information. A great benefit of SAR robots, in particular drones, is that they make it possible to collect large amounts of information, including information that was otherwise inaccessible. However, this introduces a new dilemma. Rescue workers have limited momentary cognitive capacities, and the large quantity of information, the ad-hoc nature of the operation, and the limited time of rescue workers in emergency situations make it impossible for them to inspect all the information. In order to use the information collected by robots, it needs to be automatically processed into more manageable pieces of information, e.g. by aggregating or filtering data. There is a risk that in this process relevant information is lost. Thus, on the one hand, processing information can improve performance, but on the other hand, it can also cause rescue workers to miss relevant information, which they would have noticed when not relying on SAR robots.

An example of data processing is to use images collected by drones for automated victim detection. Such technology may increase performance, but could also lead to failing detection or false positives. This yields dilemma #4: *to what extent should*

information collected by robots be processed to make it more digestible at the risk of losing information?

False expectations about robot capabilities. Stakeholders may not be able to make appropriate judgments regarding the capabilities and limitations of rescue robots, which can lead to two potential risks. On the one hand, stakeholders may overestimate the capabilities of SAR robots, which may yield false hope for victims, deployment of robots for tasks for which they are not suitable, and unjustified reliance on their performance, e.g. expectations that a robot will infallibly detect victims. On the other hand, stakeholders may underestimate a robot's capabilities, which can lead to unnecessary worries, and robots not being used to their fullest. Adequate training may contribute to more realistic expectations, thus partially solving this problem, but this may not necessarily be accessible for all stakeholders. This entails dilemma #5: *should one deploy robots, knowing that this may raise false expectations and runs the risk of degraded performance?*

Loss of privacy. The use of robots generally entails an increase in information gathering, which can potentially lead to privacy loss. This may concern personal information of rescue workers, e.g. their physical and mental stress levels, or victims, e.g. (images of) their physical condition. It can also apply to inhabitants of a disaster area, e.g. when drones collect images of their living area. If a search and rescue operation is performed in a public or semi-public building, robots may encounter personal information about employees or maybe even classified information.

Potential loss of privacy because of robot use does not necessarily result in an ethical dilemma, as it can be argued that due to the critical nature of a SAR mission, the benefits of collecting information largely outweigh the harm it may cause. This presumes, however, that the information is handled carefully, i.e. it should stay within professional rescue organizations, and only be used for SAR purposes. Because the robot-assisted SAR typically happens in a time-critical, data-rich, high-stakes and possibly quite chaotic environment, particular care regarding privacy is appropriate.

Responsibility assignment problems. Responsibility assignment problems can apply to both moral and legal responsibility, where moral responsibility concerns the question 'Who is to blame when things go wrong?' and legal responsibility 'Who is accountable when things go wrong?' Such problems can arise when robots act independently, i.e. without human supervision. If the robot malfunctions, makes a mistake or causes harm, it may be unclear who is responsible for the damage caused: the operator, the programmer, the manufacturer or the robot itself. Responsibility assignment problems become particularly complicated when the robot has (partial) autonomy, self-learning capabilities, or is capable of making choices that were not explicitly programmed. As such, dilemma #6 is the following: *should one deploy robots that may yield responsibility assignment problems?*

7.5 Conclusion

In this paper we described the results of three Value Assessment workshops with rescue workers. We believe that the workshops provided an effective way to obtain insight in the main values and value tensions around SAR robots, and that it was particularly useful to address the perspective of not only direct but also indirect stakeholders. In future Value Assessment workshops, it would be beneficial to also involve indirect stakeholders and directly ask them for their perspective.

Based on the workshop results, we identified a list of key ethical concerns and dilemmas in the robot-assisted SAR domain. As future SAR missions will most likely involve more and more advanced robot technology, we consider it prudent to address these ethical concerns. Particularly because the SAR domain incorporates—quite literally—matters of life and death, addressing these issues are relevant and timely.

It is beyond the scope of this paper to provide actual solutions for the raised dilemmas. As such, they set a research agenda highlighting areas in need of further examination. Many of the dilemmas involve a trade-off between benefits due to using a robot versus increased risk of a particular negative outcome. In order to make considerate choices in such situations, insights and tools enabling appropriate risk assessments are needed. For instance, what is the change that a robot will stop working or cause a building to collapse? Which factors influence these risks? Currently, such estimates are often not very accurate or even impossible to make. Thus, next steps would be to make more accurate estimates of different risks associated to robot use in SAR missions.

The list of concerns and dilemmas presented in this paper is by no means intended to be comprehensive. However, we do believe that they address some of the main ethical questions in the robot-assisted SAR domain. This paper thus aims to foster discussions on roboethics in general and ethics of robot-assisted SAR in particular, and contribute to the development of the SAR domain as a whole.

Acknowledgements This work is funded by the EU FP7 TRADR project (grant no. 60963).

References

Arkin R (2009) Governing lethal behavior in autonomous robots. CRC Press

Birk A, Carpin S (2006) Rescue robotics—a crucial milestone on the road to autonomous systems. Adv Robot 20(5):595–605

Bos J, Mol E, Visser B, Frings-Dresen MH (2004) The physical demands upon (dutch) fire-fighters in relation to the maximum acceptable energetic workload. Ergonomics 47(4):446–460

Butter M, Rensma A, Boxsel Jv, Kalisingh S, Schoone M, Leis M, Gelderblom G, Cremers G, Wilt Md, Kortekaas W, et al (2008) Robotics for healthcare: final report

Chang CM, Lee LC, Connor KM, Davidson JR, Jeffries K, Lai TJ (2003) Posttraumatic distress and coping strategies among rescue workers after an earthquake. J Nervous Ment Dis 191(6):391–398

Decker M (2008) Caregiving robots and ethical reflection: the perspective of interdisciplinary technology assessment. Ai Soc 22(3):315–330

Docherty BL (2012) Losing Humanity: the case against killer robots. Human Rights Watch
Fincannon T, Barnes LE, Murphy RR, Riddle DL (2004) Evidence of the need for social intelligence in rescue robots. In: IEEE/RSJ international conference on Intelligent Robots and Systems (IROS), vol 2. IEEE, pp 1089–1095
Friedman B, Hendry D (2012) The envisioning cards: a toolkit for catalyzing humanistic and technical imaginations. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, pp 1145–1148
Friedman B, Kahn PH Jr, Borning A, Huldtgren A (2013) Value sensitive design and information systems. In: Early engagement and new technologies: opening up the laboratory. Springer, pp 55–95
Fullerton CS, McCarroll JE, Ursano RJ, Wright KM (1992) Psychological responses of rescue workers: fire fighters and trauma. Am J Orthopsychiatry 62(3):371
Geale SK (2012) The ethics of disaster management. Disaster Prev Manage Int J 21(4):445–462
Harbers M, Detweiler C, Neerincx MA (2015) Embedding stakeholder values in the requirements engineering process. In: Requirements engineering: foundation for software quality. Springer, pp 318–332
Ingram B, Jones D, Lewis A, Richards M, Rich C, Schachterle L (2010) A code of ethics for robotics engineers. In: 2010 5th ACM/IEEE international conference on Human-Robot Interaction (HRI), pp 103–104
Kruijff GJM, Janíček M (2011) Using doctrines for human-robot collaboration to guide ethical behavior. In: AAAI fall symposium: robot-human teamwork in dynamic adverse environment
Kruijff-Korbayová I, Colas F, Gianni M, Pirri F, de Greeff J, Hindriks K, Neerincx M, Ögren P, Svoboda T, Worst R (2015) TRADR project: long-term human-robot teaming for robot assisted disaster response. KI - Künstliche Intelligenz, pp 1–9
Łichocki P, Billard A, Kahn PH Jr (2011) The ethical landscape of robotics. IEEE Robot Autom Mag 18(1):39–50
Lin P, Abney K, Bekey GA (2011) Robot ethics: the ethical and social implications of robotics. MIT Press
Lin P, Bekey GA, Abney K (2009) Robots in war: issues of risk and ethics
Lucas GR Jr (2011) Industrial challenges of military robotics. J Militar Ethics 10(4):274–295
Malle BF, Scheutz M, Arnold T, Voiklis J, Cusimano C (2015) Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. In: Proceedings of the tenth annual ACM/IEEE international conference on Human-Robot Interaction, HRI '15. ACM, New York, NY, USA, pp 117–124
Miller JK, Friedman B, Jancke G, Gill B (2007) Value tensions in design: the value sensitive design, development, and appropriation of a corporation's groupware system. In: Proceedings of the 2007 international ACM conference on supporting group work. ACM, pp 281–290
Murphy RR (2004) Human-robot interaction in rescue robotics. IEEE Trans Syst Man Cybern Part C 34(2):138–153
Murphy RR (2014) Disaster robotics. MIT Press
Murphy RR, Woods DD (2009) Beyond asimov: the three laws of responsible robotics. IEEE Intell Syst 24(4):14–20
Murphy RR, Dreger KL, Newsome S, Rodocker J, Slaughter B, Smith R, Steimle E, Kimura T, Makabe K, Kon K, Mizumoto H, Hatayama M, Matsuno F, Tadokoro S, Kawase O (2012) Marine heterogeneous multirobot systems at the great eastern japan tsunami recovery. J Field Robot 29(5):819–831
Noorman M, Johnson D (2014) Negotiating autonomy and responsibility in military robots. Ethics Inf Technol 16(1):51–62
Okada Y, Nagatani K, Yoshida K, Tadokoro S, Yoshida T, Koyanagi E (2011) Shared autonomy system for tracked vehicles on rough terrain based on continuous three-dimensional terrain scanning. J Field Robot 28(6):875–893

Prieur M (2012) Council of Europe. European and Mediterranean Major Hazards Agreement (EUR-OPA). http://www.preventionweb.net/english/professional/publications/v.php?id=26384. Accessed 17 June 2015

Riek LD, Howard D (2014) A code of ethics for the human-robot interaction profession. Proceedings of We Robot

Sandin P (2009) Firefighting ethics: principlism for burning issues. Ethical Perspect 16(2):225–251

Sharkey A, Sharkey N (2012) Granny and the robots: ethical issues in robot care for the elderly. Ethics Inf Technol 14(1):27–40

Sparrow R (2007) Killer robots. J Appl Philos 24(1):62–77

Veruggio G (2006) The euron roboethics roadmap. In: 2006 6th IEEE-RAS international conference on humanoid robots. IEEE, pp 612–617

van Wynsberghe A (2013) Designing robots for care: care centered value-sensitive design. Sci Eng Ethics 19(2):407–433

Wallach W, Allen C (2008) Moral machines: teaching robots right from wrong. Oxford University Press

Zuzánek P, Zimmermann K, Hlavác V (2014) Accepted autonomy for search and rescue robotics. In: Modelling and simulation for autonomous systems: first international workshop, MESAS 2014, Rome, Italy, 5–6 May 2014, Revised Selected Papers, vol 8906. Springer, p 231

Chapter 8
Reporting Robot Ethics for Children-Robot Studies in Contemporary Peer Reviewed Papers

M. Kyriakidou, K. Padda and L. Parry

Abstract How are robot ethics described in peer-reviewed papers for children-robot studies? Do publications refer to robot ethics such as: (a) gaining children's assent, (b) providing a robot's description prior to data collection, (c) having a robot exposure phase before data collection and (d) informing children about a robot's semi-autonomy or not? A total of 27 peer-reviewed papers with an average impact factor of 1.8 were analysed. 63 % of the studies did not state any ethical procedures followed. In eight studies children gave their assent for the experiment; six studies described the robot to children prior to data collection; two studies provided a robot exposure phase prior to data collection and one study informed children that robots are operated machines. The outcomes indicate problematic applications of robot ethics in peer-reviewed journals and the necessity for the publishing industry to consider stricter actions on this aspect of a publication.

Keywords Children-robot interaction · Robot ethics · Children's assent · Ethics

8.1 Introduction

The paper's aim is to explore how robot ethics in children-robot studies are described in contemporary peer-reviewed papers.

M. Kyriakidou (✉) · K. Padda · L. Parry
Department of Psychology and Behavioural Science, Coventry University, Coventry, UK
e-mail: Marilena.kyriakidou@coventry.ac.uk

M.I. Aldinhas Ferreira et al. (eds.), *A World with Robots*,
Intelligent Systems, Control and Automation: Science and Engineering 84,
DOI 10.1007/978-3-319-46667-5_8

8.1.1 Ethics Guidelines in the Publishing Industry

There are nine main publishing companies that have presented most of the studies with children-robot interactions so far. These are: (1) American Psychological Association, (2) Elsevier, (3) IEEE (Institute of Electrical and Electronics Engineers), (4) IOS press, (5) John Benjamin's publishing company, (6) PLOS, (7) Springer, (8) Taylor & Francis, and (9) Wiley. These nine publishing companies rely on very similar documents for their ethical guidelines [e.g. American Psychological Association (2010)]. The most widely used ethical guideline is from COPE (Committee on Publication Ethics) and is used by 67 % of the publishing industry Committee on Publication Ethics (2011). Among other ethical procedures, this guideline advises authors to refer to where they have gained their ethical permission to conduct a study from and relevant ethical applications related with their study (e.g. participants' consent, informing participants on their right to withdraw). However, there are no special references in any ethical guidelines on robot ethics.

It is supported Kyriakidou (2014) that studies with robots should include special ethical measures especially when addressing children as participants. The reasons for having special ethical measures are the need for ensuring reliable data is obtained from children during their interaction with robots as well as aiming to decrease the possibilities of traumatising children.

8.1.2 Robot Ethics for Children

These special ethical measures are referred to here as robot ethics. Robot ethics are distinguished from the 'procedures due to the presence of robots within a study. Robot ethics include: (a) the importance of gaining children's assent, (b) providing a description of the robot to children prior to any experiment that will include a robot, (c) a robot exposure phase prior to the main data collection and (d) informing children about robots semi-autonomy or operation (Kyriakidou 2014).

With regard to children's assent, the UN Convention of Children Rights UNICEF (1989) is an international law stating that if children are able to understand a situation in which they are about to become involved, then children have the right to share their thoughts on that situation. Moreover, legislation [e.g. Her Majesty's Stationery Office (1969) in the UK] and ethical guidelines advise that children's assent should be obtained prior to any laboratory or field studies [e.g. Prout (2002)].

The description of a robot prior to an experiment is aimed to help prepare children for the fact that they are about to face a robot rather than immediately bringing a robot in front of a child. Most children are familiar with the robots they have seen on television and in movies, documentaries and books. Therefore, there must be a variety of perceptions of robots between children. This description should be followed by exposing children to robots for a while before the main data collection. Some children could be afraid of robots and this description can help researchers to

identify children who, for any reason, do not wish to participate in a study with a robot. Utilising the description and exposure process, the child could be gradually introduced to the robot instead of being exposed directly to it during the study's data collection. Such a smooth introduction can help to avoid any possible unpleasant discomfort for children (Kyriakidou 2014).

Explanations to children about the robot's autonomy, before or after the completion of a study demonstrate that the robot was semi-autonomous or fully operated by an operator. Ethical issues concerning the interaction between children and robots include the child becoming too attached to a robot resulting in the child being incapable of understanding the nature of a robot. This is related to 'attachment theory' as discussed by Sharkey and Sharkey (2010) that raises questions as to whether children are capable of developing an ethically problematic preference for robots in comparison to their caregivers Sharkey and Sharkey (2010). Therefore, explaining a robot's operational nature can reduce the possibilities of children developing fake relationships with the robot used in the study.

These four features, children's assent, robot's description, robot's exposure and explanation of robot's semi-autonomy are important for researchers to respect in order to ensure children's rights and wellbeing. In our knowledge, there were no studies investigating if and how these four robot ethics features are presented into peer-reviewed papers. With no references on robot ethics within the nine main publishing companies' guidelines, we were expecting a minority of publications to refer to robot ethics.

The present study focuses on how robot ethics are reported in published research and not on how research itself is applying robot ethics. It is assume that studies' methodologies described in published manuscripts will have reflected the actual studies' methodologies applied. Therefore, studies that applied robot ethics will have reported this in a publication. The study's research questions were:

Research question 1: Whether peer-reviewed papers refer to 'classical' ethics and robot ethics in their articles.

Research question 2: If yes, how the papers described robot ethics in terms of children's assent, robot description, robot exposure prior to the experiment and informing children about the robot's semi-autonomy.

8.2 Methods

8.2.1 Procedure, Sampling, Analyses

Articles were identified from the Lanchester Library search engines via Coventry University. This enabled us going through the library's online catalogue which covers 8000 electronic journals. They are however, journals relevant with robotics not enlisted in the library database that we did not have access to. This minimized our

sample into papers available in Lanchester Library. They were five keywords used to detect the papers included: 'children', 'robots', 'advanced technology', 'children and advanced technology' and 'children-robots interaction'. The search engine results were then filtered into peer-reviewed publications. Articles had to have children as participants for experiments that involved robots in a direct (e.g. face to face) or indirect (e.g. via a TV link) interaction.

All 27 articles that fulfilled these criteria: (a) published after a peer-reviewed process, (b) had children as participants and (c) included a robot were added into Statistical Package for the Social Sciences (SPSS).

Descriptive statistics were produced to address a series of questions of interest presented in the following section.

8.2.2 Sample

The sample included 27 peer-reviewed papers published in journals and conferences. Studies used in the sample are presented in the appendix section. Table 8.1 below shows the countries in which the experiments were conducted. Almost half of the studies took place in the United States. The impact factor of the journals varied from 0 to 4.235 with an average of 1.8 (SD = 1.25). The first year of publication was in 2004 with the latest being in 2014, the mode of years of publication was 2013 (representing 33.3 % of all years of publication). The number of child participants ranged from one (in one study) to 228, with an average of 42.11 participants (SD = 55.53). The youngest age group of the participants in the studies was 3 years old and the oldest 18 years old.

Robots used in the studies included: (a) therapeutic robots for children with autism (such as IROMEC, KASPAR, Paro, Probo), (b) robots for research purposes focusing on human-robot interactions such as (custom made robots, ISAC, Pioneer 3-DX,

Table 8.1 Countries where the experiments were conducted

Countries	Frequency	Percentage (%)
Canada	3	11
England	1	4
France	1	4
Israel	1	4
Japan	2	7
Netherlands	1	4
Pakistan	1	4
Romania	3	11
Switzerland	2	7
United States	12	44
Total	27	100

Qrio, Robovie, Robonova, Nao), (c) medical robots (such as da Vinci surgical robot) and (d) robots marketed as entertainment products (such as AIBO, GIPY-1, I-Cybie, iCat, Pleo, Robonova, Rovobot, Nao). There were five studies in which it was not clear what kind of robot was used. Most of the studies (12 out of 27, 44.4 %) used humanoid robots followed by six animal-like and five machinery-like robots.

Studies research questions were disseminated in four main categories. First, studies' focusing on children's learning attitudes in human-robot interactions (e.g. how pre-schoolers learn words during interaction with robots compare with human interaction?). Second, studies on how children value robots in an act (e.g. how they value robot interviewers compare with human interviewers?). Third, studies on how children interact with robots (e.g. can children interact with robots as they interact with other children?). Fourth, studies on how children can be helped by robots (e.g. how children with physical disabilities can be helped by interacting with advanced technology?).

8.3 Results

Classical ethical procedures followed during the experiments were described in 10 of the 27 articles (37 %) with the remaining 17 (63 %) papers omitting any references to ethics. The ethics section of these studies was brief and was not more than 52 words for each paper. From the 10 papers that included an ethics section all mentioned that they received an informed consent from the children's guardians. Two of these 10 papers stated that their study was approved by their academic institution. Children's assent was elicited by eight of these 10 studies. Children assent is considered both a 'classical' ethical procedure and a robot ethical procedure.

Robot ethical procedures related with robot's description prior main data collection were presented in six studies. In these studies, authors stated that they had provided a description or a brief introduction to the children about the robot prior to the main data collection. These robot descriptions were described briefly in the papers with no more than 21 words. For example, a paper stated 'the child was instructed that s/he was going to learn some words from the agents (robot)' (Moriguchi et al. 2011) and another paper reported 'In this introductory session we provided information on the nature of the robot (KASPAR was explicitly introduced as a robot)' (Wood et al. 2013).

Robot ethical procedures related with children being exposed to the robot prior main data collection where described in two studies. Robot exposure was described in both studies as 'habituation phase lasted 5 min and Robonova was introduced to each child. The experimenter presented the robot by telling its name and describing the robot components: the eyes, the arms and the legs. Then the experimenter encouraged the child to touch the robot' (Cristina et al. 2013) and 'The children were briefly given a group introduction to both KASPAR and the human interviewer at the school one day before the interviews commenced' (Wood et al. 2013).

There was only one study that informed children about the robot's autonomy or semi-autonomy (which is considered a robot ethical procedure). This study (Somanader et al. 2011) referred to the robot's operation because it was part of its experimental design. Children in this study were divided into two conditions; one condition they were informed that the robot was controlled by an operator and the second condition they did not get any instruction on this topic.

8.4 Discussion

The papers' ethical sections (both for classical and robot ethics) were the most underestimated part of the publication with six out of every ten studies omitting this part and the minority who included it, referred to it very briefly.

Stating that informed consent was received from children's guardians was the only ethical measure mentioned by all 10 studies that referred to ethical measures. However, this is a standard 'classical' ethical process and is not considered a special ethical practice characterised as robot ethics.

When it comes to robot ethics, only eight studies reported receiving children's assent. Failing to gain children's assent prior any laboratory or field study, especially from older children (e.g. teenagers) who understand the situation they are about to become involved in, is a violation of important international legislation [e.g. UNICEF (1989)]. Such procedures should not be underestimated by authors and publishers.

Moreover, only six studies described the robot to children prior to the conduct of the experiment and only two studies provided robot exposure prior to the main data collection. Description of the robot and robot exposure are considered two important robot ethical procedurals. There were no clear indications from the papers analysed as to whether any children did not want to participate in a study due to the robot's presence, or whether any children were uncomfortable with the robot interaction. Some children may feel discomfort in front of robots and a slow exposure to robots can reduce upsetting children as well as identify children that may not wish to participate in a study with a robot (Kyriakidou 2014). The researchers' choice not to gradually introduce the robot to children can increase the possibilities of some children feeling uneasy during the study. Overlooking these special robot ethics should be taken into consideration prior to any publication by editors.

Only one study mentioned that the robot was operated by an operator, but this was part of the study's methodological design. Consequently, no authors mentioned that they had informed the child prior to or after the experiment that the robot they interacted with was operated. This is a challenging ethical measure and its underestimation can be characterised as deception. Misleading child participants during experimental designs must be conducted under special ethical consent and most ethical committees would rarely allow such acts. However, none of the sample publications informed children about the robot's operational nature and this leaves unanswered questions on how children understood their relationship with the robot after the completion of the experiment. Researchers should have a duty to explain the robots operational nature

to children after the experiment and if necessary show the operation equipment to the children.

The main limitation of the present study is that there is a possibility that researchers may have applied robot ethics but did not summarise them in the published manuscript. For example, the authors may have summarised the robot ethical procedures in a letter to the editor prior to any publication. Therefore, they are not included in the published paper. Future studies should ask the authors directly if they followed such ethical procedures or expand the data collection to include letters to the editors.

Ethical guidelines in the publishing industry omit references to robot ethics. As a result researchers do not state any robot ethics procedures when conducting children-robot studies. The outcomes indicate problematic applications of reporting robot ethics in peer-reviewed journals and the necessity for the journals industry to consider stricter action on this aspect of publication.

Appendix

Beran, T.N., Ramirez-Serrano, A. (2010) 'Do children perceive robots as alive?' Conference paper [online] 137–138. Available from http://dl.acm.org/citation.cfm?doid=1734454.1734511 [26 May 2015]

Beran, T. (2010). Robots, children, and helping: Do children help a robot in need? Human-Robot Interaction (HRI), 5th ACM/IEEE International Conference, March 2010.

Blanson Henkemans, O.A., Bierman, B.P.B., Janssen, J., Neerincx, M.A., Looije, R., Van Der Bosch, H., Van Der Giessen, J.A.M. (2013) 'Using a robot to personalise health education for children with diabetes type 1: A pilot study' Patient Education and Counselling, 92 (2), 174–181.

Cristina, P., Anreea, C., Sebastian, P., Andreea, P., Ramona, S., Bram, V., Dniel, D. (2013). Imitation and social behaviors of children with ASD in Interaction with Robonova. A series of single case experiments. Transylvania Journal of Psychology, 14 (1), 71.

Flannery, L.P., Bers, M.U. (2013) 'Lets Dance the Robot Hokey-Pokey!: Children Programming Approaches and Achievement throughout Early Cognitive Development' Journal of Research on Technology in Education, 46 (1), 81–101.

Fridin, M. (2014) 'Kindergarten social assistive robot: First meeting and ethical issues' Computers in human Behavior, 30, 262–272.

Giannopulu, L., Pradel, G. (2010) 'Multimodal interactions in free game play of children with autism and a mobile toy robot' NeuroRehabilitation, 27 (4), 305–311.

Hartwich, J., Tyagi, S., Margaron, F., Oitcica, C., Teasley, J., Lanning, D. (2012) 'Robot-Assisted Thoracoscopic Thymectomy for Treating Myasthenia Gravis in Children' Journal Of Laparoendoscopic & Advanced Surgical Techniques, 22 (9), 925–929.

Jipson, J.L, Gelman, S.A. (2007) 'Robots and Rodents: Children Inferences About Living and Nonliving Kinds' Child Development, 78 (6), 1675–1688.

Kahn, P.H., Kanda, T., Ishiguro, H., Freier, N.G., Severson, R.L., Gill, B.T., Ruckert, J.H., Shen, S. (2012) "Robovie, Youll Have to Go into the Closet Now": Children's Social and Moral Relationships With a Humanoid Robot' Developmental Psychology, 48 (2), 303–314.

Kanda, T., Hirano, T., Eaton, D., Ishiguro, H. (2004) 'Interactive robots as social partners and peer tutors for children: A field trial' Human-computer interaction, 19 (1), 61–84.

Kim, E., Berkovits, L., Bernier, E., Leyzberg, D., Shic, F., Rhea, P., Scassellati, B. (2013) Journal of Autism and Developmental Disorders, 43 (5), 1038–49.

Ladenheim, B., Altenburger, P., Cardinal, R., Monterroso, L., Dierks, T., Mast, J., Krebs, H.I. (2012) 'The effect of random or sequential presentation of targets during robot-assisted therapy on children.' International Journal of Neurorehabilitation, 33 (1), 25–31.

Lehmann, H., Iacono, I., Dautenhahn, K., Marti, P., Robins, B. (2014). Robot companions for children with down syndrome. Interaction Studies, 15 (1), 99–112.

Moriguchi, Y., Kanda, T., Ishiguro, H., Shimada, Y., Itakura, S. (2011) 'Can young children learn words from a robot?' Interaction Studies, 12 (1), 107–118.

Okita, S.Y. (2013) 'Self Others Perspective Taking: The Use of Therapeutic Robot Companions as Social Agents for Reducing Pain and Anxiety in Pediatric Patients' Cyberpsychology behavior and social networking, 16 (6), 436–441.

Poletz, L., Encarnação, P., Adams, K., Cook, A. (2010) 'Robot skills and cognitive performance of preschool children' Technology and Disability, 22, 117–126.

Prazak, B., Kronreif, G., Hochgatterer, A., Furst, M. (2004) 'A toy robot for physically disabled children' Technology and Disability, 16 (3), 131–136.

Ribi, F.N., Yokoyama, A., Turner, D.C. (2008) 'Comparison of Children's Behavior toward Sony's Robotic Dog AIBO and a Real Dog: A Pilot Study' Anthrozoos: A Multidisciplinary Journal of The Interactions of People & Animal, 21 (3), 245–256.

Saylor, M.M., Somanader, M., Levin, D.T., Kawamura, K. (2010) 'How do young children deal with hybrids of living and non-living things: The case of humanoid robots' British Journal of Developmental Psychology,28 (4), 835–851.

Schoepflin, Z.R., Chen, X., Ragonesi, C.B., Galloway, J.C., Agrawal, S.K. (2011) 'Design of a novel mobility device controlled by the feet motion of a standing child: a feasibility study' Journal of Medical and Biological Engineering, 49 (10), 1225–1231.

Schuler, T.A., Müller, R., Van Hedel, H.J.A. (2013) 'Leg surface electromyography patterns in children with neuro-orthopedic disorders walking on a treadmill unassisted and assisted by a robot with and without encouragement' Journal of Neuroengineering and Rehabilitation, 10, 78.

Shahid, S., Krahmer, E., Swerts, M. (2014) Child-robot interaction across cultures: How does playing a game with a social robot compare to playing a game alone or with a friend? Computers in Human Behavior, 40, 86–100.

Somanader, M.C., Saylor, M.M., Levin, D.T. (2011) 'Remote control and children's understanding of robots' Journal of experimental child psychology, 109 (2), 239–247.

Tapas, A., Peca, A., Aly, A., Pop, C., Jisa, L., Pintea, S., Rusu, A., Alina, S., David, D.O. (2012) 'Children with autism social engagement in interaction with Nao, an imitative robot' Interaction Studies [online] 13 (3), 315–347.

Vanderborght, B., Simut, R., Saldien, J., Pop, C., Rusu, A.S., Pintea, S., Lefeber, D., David, D.O. (2012) 'Using the social robot Probo as a social story telling agent for children with ASD' Interaction Studies, 13 (3), 348–372.

Wood, L.J., Dautenhahn, K., Rainer, A., Robins, B., Lehmann, H., Syrdal, D.S. (2013) 'Robot-Mediated Interviews—How Effective Is a Humanoid Robot as a Tool for Interviewing Young Children?' PLoS One, 8 (3).

References

American Psychological Association (2010) Ethical Principles of Psychologists and Code of Conduct

Committee on Publication Ethics (2011) Code of Conduct

Her Majesty's Stationery Office (1969) Family Law Reform Act

Cristina P, Andreea P, Sebastian P, Ramona S, Bram W, Deviel D (2013) Imitation and social behaviors of chidren with ASD in interaction with Robonova. A series of single case experiments. Transylvanian. J Psychol 14(1):71

Kyriakidou M (2014) Discussing robot crime interviewers for children's forensic testimonies: a relatively new field for investigation. AI Soc. doi:10.1007/s00146-014-0566-3

Moriguchi Y, Kanda T, Ishiguro H, Shimada Y, Itajura S (2011) Can young children learn words from a robot? Interact Stud 12(1):107–118

Prout A (2002) Researching children as social actors. Child Soc 16(2):67–76

Sharkey N, Sharkey A (2010) The crying shame of robot nannies: an ethical appraisal. J Interact Stud 11:161–190

Somanader M, Saylor M, Levin D (2011) Remote control and children's understanding of robots. J Exp Child Psychol 109:239–247

UNICEF (1989) UN Convention on the right of the child

Wood L, Dautenhahn K, Rainer A, Robins B, Lehmann H, Syrdal D (2013) Robot-mediated interviews: how effective is a humanoid robot as a tool for interviewing young children? In: PLOS, International Conference on Social Robotics, 2013

Chapter 9
A Typology of Liability Rules for Robot Harms

Sjur Dyrkolbotn

Abstract This paper considers non-contractual liability for harms caused by (artificially) intelligent systems. It provides a typology of different ways to approach the liability issue, exemplified by some new technologies that have been, or are about to be, introduced into human society. The paper argues that the traditional robot-as-tool perspective should be maintained, but warns that this might not be possible unless we develop corresponding technologies for efficient responsibility tracking. Specifically, new techniques need to be developed, at the intersection between computer science and law, to support reasoning about the liability implications when autonomous technologies interact with their environment and cause harms.

Keywords Non-contractual liability for intelligent systems · Individual versus collective agents · Liability regime · Prohibition · Immunity

9.1 Introduction

The issue of non-contractual liability for intelligent systems is becoming an important topic for legal scholars and policy makers.[1] As pointed out in the final report of the EU-funded RoboLaw project, liability rules also have important *ex ante* effects, as they provide incentives to technology providers, investors and end users.[2] A well-designed liability regime should not be so strict as to discourage investment, but

[1] See, generally, Cerka et al. (2015); Hallevy (2015); Vladeck (2014); Allain (2013); Gurney (2013).
[2] See (Palmerini et al., 2014, 175).

S. Dyrkolbotn (✉)
Department of Philosophy and Religious Studies, Utrecht University,
Utrecht, The Netherlands
e-mail: s.k.dyrkolbotn@uu.nl

M.I. Aldinhas Ferreira et al. (eds.), *A World with Robots*,
Intelligent Systems, Control and Automation: Science and Engineering 84,
DOI 10.1007/978-3-319-46667-5_9

should be just strict enough to ensure that a suitable proportion of the investment is directed at safety and upholding ethical standards.

In this paper, I present six kinds of liability regimes for AI systems, motivated by examples and anchored in previous scholarship. I argue that a conservative robot-as-tool approach, based on products liability rules, is the best way forward. At the same time, the challenge of applying existing rules to highly complex technologies points to a new avenue of research, at the intersection between law, philosophy and computer science. To apply products liability rules to new AI technologies, we need effective heuristics for tracing responsibilities through causal chains that involve both artificial and human agency. This paper argues that more research needs to be directed at developing systems that can aid us in this task, to ensure that our liability rules function as intended, both as incentives to manufacturers and end-users and as a source of justice for the victims of robot harms.

9.2 The Conservative Approach

The conservative approach to liability for intelligent systems is to apply established principles of products liability. This means that a manufacturer will usually only be held liable for harms that can be attributed to a "defect" of their product.[3] As a result, liability disputes will tend to revolve around inspecting and analysing the product in question and the sequence of events that caused harm. This will involve a detailed investigation aiming to shed light on *why* the product in question failed to live up to expectations.

The information resulting from such an investigation needs to be made accessible to lawyers, judges, and jury members (in some jurisdictions), along with information about relevant safety standards and typical expectations among end users.[4] On this basis, it must be decided whether the harm resulted from a failure on part of the manufacturers. When complex technologies are involved, this analytic mindset can make it hard for plaintiffs to come up with sufficient evidence to support their claim for damages. For a concrete example, consider the so-called *da Vinci* surgical system, a robot performing surgery while being remote-controlled by human doctors.[5] Recent liability cases arising with respect to this system show how a robot-as-product perspective risks turning technological complexity into a *de facto* defence against liability, possibly resulting in unfairness and socio-economic inefficiency.

A great number of lawsuits have been filed against Intuitive Surgical, the company delivering the *da Vinci* system.[6] However, it has proven very difficult to make products

[3] See, generally, Howells and Owen (2010).

[4] See Howells and Owen (2010, 241).

[5] The benefit of using robots to perform surgery are obvious; robot hands are steady and precise, more so than the hands of human surgeons. In addition, the *da Vinci* system makes it possible for human doctors to perform surgery from afar, so they can help people without being physically present. The system has proved successful, and so far it has been deployed to around 2500 hospitals around the world. See Kirkpatrick (2014, 14).

[6] Apparently, by October 2012 around 3000 claims had been submitted, see Moylan (2014).

liability claims stick in these cases. For instance, in the US case of *Mracek v Bryn Mawr Hospital*, the patient, Mracek, suffered erectile dysfunction and abdominal pain after having had his prostate removed by *da Vinci*.[7] The system had experienced technical problems during surgery, and Mracek therefore sued both the hospital and Intuitive Surgical. However, the case did not even proceed to trial, since summary judgement was granted in favour of the defendants.

According to the court, the expert testimony had been insufficient to back up the claim that the *da Vinci* surgical system was to blame. This was the conclusion even though *da Vinci* itself had displayed error messages during the operation and had stopped taking commands from the human operator. Importantly, the court required the plaintiff to establish *more* than just a causal link between the robot's unplanned behaviour and the harm suffered. Expert testimony also had to be provided to substantiate the claim that the system had malfunctioned. Mracek argued that this was obvious in light of the robot's error messages and its disobedience, but this argument was rejected.[8]

Clearly, the court thought about the robot as a product, not an autonomous agent. Hence, the typical question arose: was the harmful behaviour of the product the result of a defect or something else, e.g., human error or a lack of maintenance? In the absence of any concrete proof of a defect, the case had to be dismissed. In particular, the agency of *da Vinci* did not by itself count as evidence of wrongdoing. Rather, it was expected that the robot's actions would be picked apart and analysed in order to get at underlying causes that could be attributed to, or at least connected with, some sort of *human* agency.

This way of thinking is well established, but raises serious problems in cases such as *Mracek*. In practice, it might be close to impossible to deliver a technically adequate explanation of *why* a system such as *da Vinci* behaves in a certain way under a given set of circumstances. Moreover, an approach based on holding manufacturers to some "industry standard" might fail, especially when there is a lack of reliable information from independent sources.[9] In short, the demand for expert testimony can become an insurmountable hurdle for a plaintiff seeking damages. Unfortunately, this can be so even in cases when it is regarded as obvious that wrongful damage has been inflicted. Indeed, if a human surgeon suddenly stops operating while shouting "error, error", the liability question is hardly in any doubt. Moreover, there would

[7] *Mracek v Bryn Mawr Hospital*, 610 F Supp 2d 401 (ED Pa 2009), aff'd, 363 F App'x 925 (3d Cir 2010).

[8] For a more in-depth discussion of the case, see Pagallo (2013, 91–95) and Goldberg (2012).

[9] This is particularly likely to become a problem when, as in the case of *da Vinci*, the manufacturer is a *de facto* knowledge monopolist, with access to privileged information about how the system actually works. See Goldberg (2012, 249) ("Because Intuitive has a monopoly, as well as limited expert witnesses, manipulation of an industry standard by such a dominant player is likely in a situation where they would be protecting themselves in a products liability lawsuit."). This also points to the important interactions between liability rules and intellectual property rights, a connection that could provide a strong incentive for large technology firms to favour an approach based on insurance payments and new forms of strict liability. In this way, these firms might hope to avoid in-depth public scrutiny of their technology, both its merits and its shortcomings.

be little need for expert testimony, except perhaps for the purpose of sentencing in a criminal case (a plea of insanity, one imagines, might have to be considered).

The broader point illustrated by the *Mracek* case is that applying standard products liability rules to harms caused by complex technologies can result in a mismatch between what the law delivers and what people regard as justice. This could soon become politically intolerable, at which point we seem to have two choices: we either develop new principles and techniques for analysing products against established liability doctrines, or else we replace those doctrines by rules that do not require the same intensity of analysis and depth of technological understanding. If we take the latter route, what notions of liability might we use? In the following, I present a categorisation of what I take to be the main candidates.

9.3 The Moderate Approach

For liability rules that target humans (as opposed to products), a key notion used to determine if the defendant is responsible is the "duty of care". Liability is usually imposed only on defendants that caused harm negligently, relative to some context-dependent standard of reasonable behaviour. In some cases, harm is caused by a collective rather than a single human agent, in which case the law sometimes gives effect to expectations of reasonableness that are directed at the behaviour of the collective as such, while relaxing the requirement that negligence has to be proved at the individual level.[10] This collective type of liability, or enterprise liability, was traditionally quite controversial, particularly in criminal law, but today it forms an important part of the liability regime in many jurisdictions.[11]

Enterprise liability relies on a way of thinking that can also be applied to intelligent systems. Specifically, the behaviour of an intelligent machine can be approached as a component of the collective behaviour of a group of agents, e.g., the machine's owners, users, producers, and developers. It is then possible to attach legal consequences to artificial agency indirectly, by using it to *infer* something about the collective agency of the group. In the *da Vinci* case, for instance, it might be argued that the unexplained and harmful behaviour of the robot should count as evidence in favour of joint liability among the doctors, the hospital, and the technology providers.

This involves a radical element in that it enables us to infer liability from the unexplained agency of robots. However, it dovetails nicely on existing doctrines of enterprise liability. Hence, it is a moderate proposal for reform. As discussed in a recent article by Jessica Allain, it might be a well suited template for regulating possible AI diagnosticians such as *Watson*, an expert system originally developed by IBM to play Jeopardy.[12]

---

[10] Following this, the collective as such can be held liable, either proportionally to their causal contribution or else *in solidium* (joint liability). See, e.g., Wright (1992).

[11] See, e.g., Hamdani and Klement (2008).

[12] See generally Allain (2013).

If *Watson* is used to perform a diagnosis, he will not only do what the doctors tell him to do. Rather, *Watson*'s job will be to tell the doctors what *they* should do. Specifically, *Watson* will use his capacity for intelligent analysis of big data sets to provide recommendations that mightinfluence the treatment of patients. Given the complexity of the system, successfully tracking the reasons for *Watson*'s recommendations, and their exact consequences for a given patient, might be very difficult. To address this, Allain proposes a form of enterprise liability to analyse harms that appear to be caused by bad advice from *Watson*. The intention is to make it possible for courts to "analyse the team's fault without breaking into the interrelated actions of the individual actors".[13] If causation and fault can be established for the enterprise as such, the entire team interacting with Watson is to be held jointly liable.

Enterprise liability is only meant to be triggered with respect to *Watson* as a team member, not with respect to *Watson* as a product. In the latter regard, existing products liability rules will still apply. Specifically, Allain suggests that before enterprise liability can be established, a "panel of experts" must rule out that the cause of damage was "hardware failure".[14]

Allain's proposal makes elegant use of existing doctrines of collective responsibility. Still, I am not convinced that it succeeds in striking the right balance between the interests of the persons harmed and those who might come to be held liable. On the one hand, even the challenge of demonstrating fault with the surrounding enterprise can become quite complex, particularly when "hardware failure" first needs to be ruled out. On the other hand, there is a risk that blameless persons interacting with an intelligent machine will be held jointly liable because they inadvertently and unpredictably contribute to a causal chain that leads to harm.

For an autonomous technology such as *Watson*, situations might arise where human intuition tells us that no agent is to blame except *Watson* himself.[15] In such situations, enterprise liability for the team might effectively become strict liability (liability without fault) for humans. This could prove unsustainable, particularly if the liability is attached to human agents on the basis of a legal fiction of collective fault inferred from the agency of a machine, rather than some substantive argument in favour of strict liability. Based on the enterprise idea, anyone collaborating with *Watson* would seem at risk of being held liable for harms even when they have no control over, or even understanding of, *Watson*'s inner workings. This is hardly the best way to encourage doctors to trust *Watson*'s judgement, however good it might be in average cases.

More generally, the moderate approach might increase the demand for rules that attach *more* legal consequences to the robot's own agency, independently of the team. Asking for this would be an obvious response both to the worry that ruling out "hardware failure" is too difficult *and* the worry that blameless people might otherwise be held liable for the actions of machines they do not control. This observation leads me to the next category of possible approaches to liability for intelligent systems.

---

[13] Allain (2013, 1077).

[14] Allain (2013, 1075).

[15] Allain refers to his status as being that of a "quasi-legal person". See Allain (2013).

9.4 The Radical Approach

In certain situations, the law finds it expedient to rely on a notion of *absolute* liability, such that a plaintiff can succeed in a claim for damages as soon as proximate causation has been established.[16] If the defendant caused the damage, liability results, period. This form of liability is typically imposed in cases when the defendants fulfil roles that make it natural that they should bear the entire risk and cost of harm, e.g., because they are significantly better placed to prevent that harm or to repair the damage.

It can be tempting to apply such a liability concept to intelligent systems, especially when these are deployed in such a way that humans have no direct control over their behaviour. However, this asks us to overlook a crucial difference between human and intelligent systems as *causative* agents. For humans, it is common-place and unavoidable to abstract away from many of the deeper causes of specific actions. If you act a certain way to produce a certain outcome, we have no conceptual difficulty with saying that *you* caused the outcome, even though it was in fact your actions that made the difference.

For human agents, it is very uncommon to entirely dissociate agents from their actions, a move typically reserved for actions induced by severe mental illness, great distress, or (in some cases) intoxication. For intelligent systems, on the other hand, the natural starting point (so far) is the opposite. Even if such a system has agency, we are not normally content with saying that the system caused the harm. We also want to know *why* things went wrong. Was the underlying cause a hardware error, a software error, a virus, or simply an instance of inappropriate use?

These are natural questions, and answering them is crucial to preventing future harms. But the absolute liability approach asks us to hold our peace in this regard. This represents a dramatic shift in perspective, one that is also potentially problematic. Is it really appropriate to design the law of liability such that it no longer requires us to keep track of underlying causes of robot behaviours?

Some scholars argue that it is. In particular, an absolute liability regime has been proposed for a soon-to-become very important type of intelligent system: the self-driving vehicle.[17] Unlike surgical robots and AI diagnosticians, autonomous cars make decisions on their own that can cause harm directly. Moreover, they *will* eventually do so, and when they do, the harm might even be directly attributable to an unpredictable *choice* made by the car. For this reason, new absolute liability rules suggest themselves with increasing force.

In a recent article, David C. Vladeck argues that we should "construct a system of strict liability, completely uncoupled from notions of fault".[18] Importantly, the

[16] This should be contrasted with doctrines of strict liability, where other notions act as a substitute for *culpa*, e.g., the notion of a defect in products liability law. See, e.g., Calabresi and Hirschoff (1972, 1055–1056). From a conceptual and socio-legal perspective, it should also be kept in mind that causation itself might be used as a partly normative notion, see, generally, Hitchcock and Knobe (2009).

[17] See, generally, Vladeck (2014).

[18] Vladeck (2014, 146).

meaning of "fault" here is not limited to negligence, it is meant to cover also the technical reason why the damage occurred. That is, the liability rule is meant to be applicable as a default even when no design flaws or manufacturing mistakes can be demonstrated. In Vladeck's words, the rule is meant to support an "inference of liability drawn by operation of law to protect a blameless party (the person who sustained injured) by making others bear the cost".[19]

This is radical, as it rejects the robot-as-tool perspective. Specifically, the car's actions are essentially considered *irreducible*; if they give rise to a harm, they are in need of no further scrutiny.[20] Essentially, the car will be treated as an agent with no principal.[21] The important implication is that we are no longer compelled to understand why the car acts as it does, beyond judging the causal *effects*, to determine whether or not its behaviour resulted in damage that should be financially compensated. As discussed in the next section, this might well be the first step towards granting legal personhood to AI systems.

9.5 The Revolutionary Approach

Vladeck argues extensively in favour of his liability rule, pointing out that it spreads costs and ensures that the damaged party is not left unprotected as a result of the complexity of the causal chain that led to the damage.[22] However, he also notes that it goes quite far towards attributing full-blown legal agency to cars. This, indeed, is why his proposal is not simply a new application of absolute liability, but a radical conceptual shift in liability law.

Vladeck himself speculates that a system granting legal personhood to certain intelligent systems might be the natural next step.[23] This might have several practical advantages. For instance, it would be possible for intelligent systems to legally own assets, from which damages could then be sought. In this way, one could hope

[19] Vladeck (2014, 149).

[20] Vladeck argues that this is close to the legal doctrine of *res ipsa loquitur* (going as far as to state that it is merely a "restatement" of it). See Vladeck (2014, 128). I do not agree. According to Vladeck's proposal, liability will be inferred even in the *complete absence* of circumstantial/statistical evidence to suggest a design flaw—in fact, it might be inferred even in the presence of conclusive evidence suggesting that the autonomous car is a safer driver than most humans. The proposal therefore appears very different from the original doctrine of *res ipsa loquitur*, which is a principle used to *infer* negligence from circumstantial evidence. See, generally, Johnson (1997).

[21] Vladeck suggests that a vicarious liability system should be built on this basis, so that human actors can be held accountable for the agency of the cars even if they are entirely blameless. Specifically, he proposes an enterprise model involving not only the manufacturer but also the designers of the AI software and other technology suppliers, see Vladeck (2014, 148–149).

[22] Vladeck (2014, 145–149).

[23] Vladeck (2014, 150).

to allocate costs and risks of damage more smoothly, among all those that interact financially with the system (including end-users and insurance companies). In addition, fewer cases might end up in court, with risks and damages dealt with instead by new kinds of insurance arrangements and other market mechanisms that could benefit from a legal fiction whereby intelligent machines are considered economic agents in their own right.

Vladeck only notes that this is a natural next step, he does not go as far as to propose it. However, several other authors have proposed legal personhood for intelligent systems, coming to this conclusion from several different angles.[24] I think their suggestions deserve to be labelled *revolutionary*.

The proponents of the legal personhood idea might disagree with the connotations of such a label. Indeed, they typically only propose legal personhood for certain kinds of intelligent systems, and only in relation to certain kinds of rights and responsibilities, similarly to how many jurisdictions approach corporations today. More generally, what most of them share is a down-to-earth feeling that when intelligent systems *actually* start functioning as autonomous agents without clear principals, the law needs to follow suit. The argument is that the conceptual vacuum that arises needs to be filled by a suitable legal category that accurately reflects the facts of life.

However, this pragmatic perspective downplays the considerable *normative* power of granting legal personhood to machines. It might well be that intelligent systems and robots will soon *seem* to be completely autonomous and without principals, but should we *allow* this, and should we give up looking for human causes and intentions underlying their behaviour? That, to me, sounds rather like admitting that we no longer control, or even understand, the technology that surrounds us. It would perhaps be cost-effective to set up a legal framework that abstracts away from the workings of complex technologies, but would it be morally and politically appropriate?

It might be argued that a liability regime based on legal personhood for machines will be both efficient and strict. Those who suffer wrongs might have better access to compensation and this might in turn provide an incentive for technology developers to invest in safety and ethical standards. Moreover, it might increase confidence among end users and deliver better access to justice for victims of harms.

However, this effect might be achievable in the short term only, since safety standards themselves might in turn be weakened. Specifically, it seems that the legal personhood fiction would provide a strong incentive towards *quantitative* rather than *qualitative* approaches to safety and ethics. When there is little or no room to fend off liability claims once harm occurs, attention is likely to be directed at *average* cases and typical behaviour, rather than attempting to develop strategies for in-depth

[24] See, e.g., Chopra and White (2011, 189–191); Koops et al. (2010, 560–561) ("The majority view in the literature is that sooner or later, limited legal personhood with strict liability is a good solution for solving the accountability gap, particularly in contracting, and for electronic agents, this may be sooner rather than later"). For a bolder approach, arguing that notions of criminal liability should apply to artificial intelligences, see Hallevy (2013, 177–178).

understanding of abnormal events. As long as abnormalities are dealt with by efficient liability rules and there are not too many of them, it will be much more cost-effective to simply pay the damages, without performing deep analysis every time something unexpected happens.

The traditional approach of establishing safety by performing tests is likely to be highly influential. The worry is that this might give both regulators and laypeople a false sense of security, especially when combined with an absolute liability regime that offers case-by-case relief while potentially keeping structural problems out of sight. Hence, what seems pragmatic and expedient at first sight, particularly from the point of view of technology optimists, might well backfire later on, as people gradually loose their grasp on the distinction between reality and legal fiction. If this happens, the technology pessimists are also more likely to come into power, potentially leading us to the liability rule discussed in the next section.

9.6 Prohibition

In a sense, prohibition is the strictest possible liability rule that can be applied to an artificial intelligence: prevent its development and use on pain of criminal responsibility. It might well become the preferred option for some kinds of intelligent systems. For example, consider the case of so-called "killer drones", unmanned aerial vehicles that the US military and others are already using to kill suspected enemies.[25] The drones used for this purpose are not yet fully autonomous, they are operated by human pilots. However, in a manner similar to the *da Vinci* system, they make use of complex technology to receive and interpret the pilot's instructions, making it possible to provide such instructions from afar.[26]

The next step is obvious: complete automation. Some argue that taking this step is a *moral imperative*.[27] Just as autonomous cars will eventually outperform human drivers, it can be argued that autonomous killer drones will eventually outperform human pilots. In a not-too-distant future, delegating the kill decision to the machines might reduce the risk of mistakes and the number of unnecessary casualties. Letting computers kill, it can be argued, is going to save many lives.[28]

This argument in favour of killer drones has proved controversial, to say the least. While it is true that self-driving cars and killer drones rely on similar technologies, the "gut feeling" response to a killer drone is predictably more sceptical. Indeed, when discussing drones, our attention is not usually directed at how to tweak

[25] See, generally, Rosen (2011).

[26] See, e.g., Bachmann (2013).

[27] See Strawser (2010); Arkin (2010).

[28] See Arkin (2010).

existing liability rules. Rather, we discuss whether or not automated killing should be *forbidden*, as a principle of international law.[29]

The prohibition rule clearly belongs somewhere near the end of the line of possible liability principles. But it is not the final station. That honour belongs to the principle at the other extreme, discussed in the next section.

9.7 Immunity

From 2:42 pm to 2:45 pm on May 6, 2010, the Dow Jones Industrial Average dropped by more than 5 %, an unprecedented event in the history of modern finance.[30] The pricing of stocks in some individual companies displayed behaviour that was patently absurd. For instance, the price of shares in the consultancy company Accenture Plc dropped from $30 to $0.01 per share in seven seconds.[31] Luckily for the world economy, the crash—dubbed the *flash crash* by commentators—did not last very long. In fact, the subsequent surge in prices was just as abrupt and dramatic as the drop had been. By 3 pm, the Dow had recouped all its losses. In four seconds, the shares in Accenture rose from $0.01 to $39.[32]

How could this happen? The answer, so we have been told, was that a human ordered a computer to sell a large number of financial instruments without taking into account price or time, only volume.[33] As a result, a cascading effect resulted, whereby trading algorithms ordered to trade at certain thresholds were triggered on a massive scale. The computers did not realize that this behaviour soon became irrational, so they continued trading along a rapidly moving downward trajectory. This continued until circuit breakers kicked in to provide stabilization and reversal.

With respect to the notion of liability, the importance of the flash crash lies in how it demonstrates a different kind of harm than that considered in previous sections. The flash crash, in particular, was the result of cascading effects on a *network*. On a financial market, there is a relatively long distance from the network to the underlying physical reality of the assets that are traded, making it possible to limit the damage using circuit breakers. But this only works to some extent, and in the short term.

For other kinds of networks, more intimately connected to physical reality, circuit breakers that achieve this may not be available at all, they may fail, or—as is often the case in major events—they might make the problem worse.[34] In some cases, the

[29] See, generally, Sharkey (2010).

[30] See generally Keller (2012).

[31] Keller (2012).

[32] Keller (2012).

[33] See Keller (2012).

[34] Two famous examples are the Italy 2003 electricity outage that left virtually all of Italy without electricity, and the 2011 Southwest blackout, which affected more than 7 million people in the US. See, generally, Buldyrev et al. (2010).

initial event triggering the cascade might be attributable to specific actions, perhaps even an act of negligence.

But this hardly means that the cascade is "understood". In the case of the flash crash, one could point the finger at the ill-advised behaviour of the trader that instructed the computer that initially set the crash in motion. Indeed, this is the strategy now pursued by the US federal government, who recently had an individual trader arrested in the UK, basically accused of causing the flash crash by giving algorithms the "wrong impression" (placing fake orders and then retracting them) about the likelihood of future price movements.[35]

However, many argue that the blame lies with the computers themselves, and with those who use them exactly as they are meant to be used. Some investors have taken this reasoning to its logical conclusion by demanding compensation for damages caused by algorithmic trading in the US.[36] The only problem is that traders, as a group, enjoy *immunity* for the damage they cause by using standard trading algorithms. Indeed, the law as it stands is not prepared to apply enterprise liability to all those who use algorithmic traders, for damages caused by their combined effect on computer-driven markets. As a surrogate, tort actions have been brought against the market providers, for failing to deliver a fair playing field. However, no such claims have been given a full hearing in a US court, further indicating that we lack a legal basis for holding anyone accountable at the network level.[37]

Since intelligent systems are so often networked, we might end up grappling with this problem for many of the harms that such systems will cause in the future. It is very difficult, or even impossible, to apply standard liability concepts to networked harms, at least without first finding new (algorithmic) ways of analysing and imposing collective responsibility. Moreover, for these kinds of harms, strict liability and legal personhood for machines would not be an effective response, since these techniques fail to address the cascading effects that blur our impression of causality in these cases.

9.8 Making the Conservative Approach Work

In my opinion, the conservative approach to intelligent systems is not only the safest option, but also in a sense the most interesting and forward-thinking. It asks us, in particular, to develop techniques for truly *understanding* the deontic workings of autonomous technology deployed in unpredictable settings. This is a formidable task, but if we devote enough attention to it, it is likely to bring human knowledge forward. This knowledge could then be built up alongside fruitful development of

[35] See Goodley (2015). In my opinion, this approach to liability for the flash crash is highly problematic, particularly as the damage created by cascades will tend to be completely out of proportion to any mistakes or even bad intentions that originally set them in motion.

[36] See Smith (2014).

[37] See Hope (2015).

intelligent technologies themselves. To achieve success in this regard, we need to direct more attention at a relatively recent research question in artificial intelligence, namely the challenge of designing algorithms for efficient *responsibility tracking*.[38]

Specifically, we need to develop tools that can help us distil legally and ethically relevant information from complex causal chains involving intelligent systems. This involves more than simply describing or mapping out what happened in computer science terms, e.g., by providing a log of computations performed. Indeed, if we require all the details, such a log would contain millions of pages. On the other hand, if we abstract away from the nature of computation and only ask for a log that maps out the human perspective on the event, the log might contain little of relevance to the question of *why* the event occurred. Neither would be helpful when applying a products liability rule to analyse a harm.

This highlights a crucial challenge, namely to arrive at representation formalisms that abstract away from irrelevant details, but keep track of those details that are relevant to uncovering underlying technical causes and their deontic status. Moreover, we need to develop abstractions and analytic tools that are sensitive to underlying legal principles, so that we get hold of the information we need to preserve the robot-as-tool perspective in legal reasoning.

In order to do this, we will have to make use of intelligent systems themselves; the reasoning tasks involved are too resource consuming for human minds, especially when they involve big data sets (which is the norm when an AI system acts in an open environment). Put simply, we need computers to help us analyse computers, and we need robot judges to help us judge robot harms. In this regard, the job of the intelligent system is to provide information that humans can use to gain understanding, not to replace human assessment. If we can harness computer technology in this way, it will help us deliver justice and communicate incentives more effectively, without giving up on the constraint that we should base our conclusions on insight into the technologies we judge. We can then avoid slippery-slope strict liability rules that assign liability without fault to people who happen to stand in legal proximity to some unpredictable intelligent system. Moreover, we can avoid the abstraction of legal personhood for machines and the potentially harmful anthropomorphic perspective on computer systems that such a legal fiction would induce.

There might still be a role for variants of strict liability, to ensure a right to compensation for damages in cases when responsibility tracking falls short of clarifying the relevant causal chains. However, if the role of strict liability becomes too great, it threatens to undermine our proposed solution. Specifically, it seems appropriate to insist that our liability regime should not become an excuse for tolerating greater and greater ignorance regarding the ethical and societal implications of new kinds of intelligent systems and their behaviour. Normative assessment based on technical insight should inform the legal system in its dealings with AI, so that the law may remain based on facts rather than fictions.

[38]The REINS project at Utrecht University, with which I am affiliated, aims to make a contribution in this regard, see Broersen (2014).

The novel research challenge is to develop analytic tools that can be used to analyse intelligent systems, to map out what they are responsible for in different contexts. This is not the same as ensuring that robots behave ethically and in accordance with the law. In fact, this second challenge is conceptually more problematic, particularly under a so-called "strong reading", whereby this calls for a complete formal specification of what it means to behave "ethically" and "lawfully" in computer realisable terms. If possible at all, this is certainly not a task that computer scientists will complete any time soon.[39] The increasing demand for "ethical" AI is therefore also a double-edged sword. Specifically, the worry is that we will eventually *change* our legal and ethical principles, to facilitate computer implementations. This, unlike many other scenarios, is a plausible vision of a world where the robots have truly outdone us; a world in which our desire for fictions concerning *their* nature causes us to change deeply rooted aspects of our own.

To avoid such a world, ethicists and lawyers should work together with computer scientists. Their aim should not be to make robots "understand" human notions, but to help us understand robots, so that we may withstand the temptation for anthropomorphic reasoning about their behaviour. Despite rumours to the contrary, it is still much easier for us to adopt the perspective of the robot, then it is for the robot to adopt the perspective of the human. As long as this remains the case, there is little doubt that the most important agents in robot law and robot ethics will all remain human for quite some time to come.

9.9 Conclusion

This paper has presented a typology of liability rules that deal with artificially intelligent systems. The proposed classification started with standard products liability rules, moved on to consider new forms of enterprise liability, then absolute liability, before reaching proposals based on granting legal personhood to intelligent systems. The paper also discussed two degenerate approaches, prohibition and immunity, both of which become likely options if we fail to come up with adequate substantive notions.

I argued that imposing absolute liability for harms caused by intelligent agents is a radical approach, more so than it might seem at first sight. This is because it proposes a deep change to the way the law approaches the agency of intelligent systems, asking the law (and us) to take this agency at "face value", similarly to how we approach human behaviour. The next step, legal personhood for intelligent systems, might then become very hard to resist.

In the end, I argued in favour of continued reliance on standard products liability rules. The main challenge associated with this approach is to develop techniques and tools for tracking liabilities through complex causal chains involving intelligent

[39] There is interesting work being carried out in this direction, but it remains rather embryonic, see, e.g., Dennis et al. (2015).

systems. In my opinion, this is a challenge we should now take on as a collective responsibility. The goal should be to ensure that human dominion is preserved, on terms that remain humane and in a way that enables us to identify the "ghost in the machine", the real human agency behind artificial intelligence.

References

Allain JA (2013) From Jeopardy! to jaundice: the medical liability implications of Dr. Watson and other artificial intelligence systems. La Law Rev 73:1049

Arkin RC (2010) The case for ethical autonomy in unmanned systems. J Mil Ethics 9(4):332–341

Bachmann SD (2013) Targeted killings: contemporary challenges, risks and opportunities. J Confl Secur Law 18(2):259

Broersen J (2014) Responsible intelligent systems. KI - Künstliche Intelligenz 28(3):209–214

Buldyrev S, Parshani R, Paul G, Stanley H, Havlin S (2010) Catastrophic cascade of failures in interdependent networks. Nature 464(7291):1025–1028

Calabresi G, Hirschoff JT (1972) Toward a test for strict liability in torts. Yale Law J 81(6):1055–1085

Cerka P, Grigiene J, Sirbikyte G (2015) Liability for damages caused by artificial intelligence. Comput Law Secur Rev 31(3):376–389

Chopra S, White LF (2011) Legal theory for autonomous artificial agents. University of Michigan Press

Dennis LA, Fisher M, Winfield AFT (2015) Towards verifiably ethical robot behaviour. AAAI-15 Workshop on AI and Ethics (to appear)

Goldberg M (2012) The robotic arm went crazy! The problem of establishing liability in a monopolized field. Rutgers Comput Technol Law J 38(2):225

Goodley S (2015) 'Flash crash trader' Navinder Singh Sarao loses bail appeal. The Guardian Available at http://www.theguardian.com. Accessed 31 Mar 2016

Gurney JK (2013) Sue my car not me: products liability and accidents involving autonomous vehicles. Univ Ill J Law Technol Policy 2013(2):247–277

Hallevy G (2013) When Robots Kill: artificial intelligence under criminal law. Northeastern University Press

Hallevy G (2015) Liability for Crimes Involving Artificial Intelligence Systems. Springer

Hamdani A, Klement A (2008) Corporate crime and deterrence. Stanford Law Rev 61(2):271–310

Hitchcock C, Knobe J (2009) Cause and norm. J Philos 106(11):587–612

Hope B (2015) Lawsuit against exchanges over 'unfair advantage' for high-frequency traders dismissed. Wall Street J. http://www.wsj.com. Accessed 31 Mar 2016

Howells G, Owen DG (2010) Products liability law in America and Europe. In: Howells G, Ramsay I, Wilhelmsson T, Kraft D (eds) Handbook of research on international consumer law. Edward Elgar Publishing, chap 9, pp 224–256

Johnson MR (1997) Rolling the "barrel" a little further: allowing res ipsa loquitur to assist in proving strict liability in tort manufacturing defects. William Mary Law Rev 38(3):1197–1255

Keller AJ (2012) Robocops: regulating high frequency trading after the flash crash of 2010. Ohio State Law J 73(6):1457

Kirkpatrick K (2014) Surgical robots deliver care more precisely. Commun ACM 57(8):14

Koops BJ, Hildebrandt M, Jaquet-Chiffelle DO (2010) Bridging the accountability gap: rights for new entities in the information society? Minn J Law Sci Technol 11:497

Moylan T (2014) Da vinci surgical robot maker reserves $67M to settle product liability claims. https://www.lexisnexis.com/legalnewsroom. Accessed 31 Mar 2016

Pagallo U (2013) The laws of robots: crimes, contracts, and torts. Springer

Palmerini E, Azzarri F, Battaglia F, Bertolini A, Carnevale A, Carpaneto J, Cavallo F, Carlo AD, Cempini M, Controzzi M, Koops BJ, Lucivero F, Mukerji N, Nocco L, Pirni A, Shah H, Salvini P, Schellekens M, Warwick K (2014) Robolaw, guidelines on regulating robotics. http://www.robolaw.eu/RoboLaw_files/documents/robolaw_d6.2_guidelinesregulatingrobotics_20140922.pdf. Accessed 31 Mar 2016

Rosen RD (2011) Drones and the U.S. courts. William Mitchell Law Rev 37:5280

Sharkey N (2010) Saying 'no!' to lethal autonomous targeting. J Mil Ethics 9(4):369–383

Smith A (2014) Fast money: the battle against the high frequency traders. The Guardian. http://www.theguardian.com. Accessed 31 Mar 2016

Strawser BJ (2010) Moral predators: the duty to employ uninhabited aerial vehicles. J Mil Ethics 9(4):342–368

Vladeck DC (2014) Machines without principals: liability rules and artificial intelligence. Wash Law Rev 89:117

Wright RW (1992) The logic and fairness of joint and several liability. Memphis State Univ Law Rev 23(1):45–84

Chapter 10
Safety and Ethical Concerns in Mixed Human-Robot Control of Vehicles

Endre E. Kadar, Anna Köszeghy and Gurvinder Singh Virk

Abstract Many robotic applications require human-like behaviour of an artificial agent and quite often include mixed (human-robot) control set-up for effective operation and to meet regulatory requirements. Developing such systems with high level of autonomy implies a good understanding and variability of human behaviour. One of the most popular areas of research in robotics with mixed control is to develop self-driving cars that are able to participate in normal traffic scenarios and acceptable under established risk management processes. In Berlin, some cars have already been licensed, but manufacturing driverless cars is more difficult than usually assumed. This is mostly because vehicular control should be human-like to avoid confusing pedestrians, passengers or other human drivers. Among the many difficulties to achieve human-like control in ensuring safety requirements are satisfied as well as being ethically acceptable, the problem of identifying and calibrating control parameters is far more complex than traditional control and systems theory alone would be able to handle. The paper provides insights into the difficulties of autonomous and mixed vehicle control and generally warns about the theoretical and ethical consequences of our limited understanding of human performance issues in car driving.

Keywords Autonomous vehicle control · Parameters invariants of human visual control · Perceptual invariants · Engineering control strategies · Human perceptual control strategies

E.E. Kadar (✉)
Department of Psychology, Keimyung University, Daegu, Korea
e-mail: Endre.kadar@port.ac.uk

E.E. Kadar · A. Köszeghy
Department of Psychology, University of Portsmouth, Portsmouth, UK

G.S. Virk
InnotecUK Ltd, Cambridge, UK
e-mail: Gurvinder.Virk@innotech.com

M.I. Aldinhas Ferreira et al. (eds.), *A World with Robots*,
Intelligent Systems, Control and Automation: Science and Engineering 84,
DOI 10.1007/978-3-319-46667-5_10

10.1 Introduction

Recently, several car-manufacturing companies are competing to produce the first commercially available autonomous cars that are able to participate in normal traffic scenarios while complying with mandatory safety regulations to prevent accidents. In particular, the primary goal of the risk assessment and risk reduction in this application is to ensure safe driving performances without risking human lives (i.e., by reducing the overall number of accidents that involve pedestrians or other vehicles with human drivers and passengers). Ethical issues are usually overlooked but it is important to highlight that the performance of the self-driving car should not cause stress in human passengers and various other participants of the traffic (i.e., the system should also be ethically acceptable). The problem of automatic vehicle control could be regarded as part of control and systems theory within which calibration is usually considered as part of the design of a complex overall control mechanism. The usual objective of a control theory-based approach is to calculate solutions for the proper corrective action by the controller that intends to move the system to a desired future (goal) state while maintaining the stability of the controlled system. For instance, let us consider an automated (or human) vehicle controller, which can navigate while controlling the direction of speed of the vehicle to arrive at a destination. The controller is taking information about constraints of the surface layout of the environment (obstacles, elevation of road, etc.) and other parameters such as the size, mass of the vehicle, friction, power of the engine, wind speed and direction, etc. Generally, these are the parameters of the system dynamics from which some are constant (e.g., size of the vehicle), while others (e.g., most of the environmental parameters) vary. The same control system could be used in various vehicles by calibrating them for differences in their parameters. The system output is the vehicle's speed and the direction of motion based on the power of the engine and the steering mechanism.

Autonomous vehicle control is not a new problem in robotics. Indeed, there are many driverless vehicles already in operation including fork-lift trucks, trains, subway and drones. In many ways, their control systems could be even safer than human control given the low level of complexity of the operational environment and the predictability of the control actions. The problem of analysing the continuously changing environment layout for prompt and correct decision-making is simple and based on redundant sensory and control systems to avoid problems typical of human operators (e.g. subject to distractions, lack of attention, sudden heart attacks, losing consciousness, etc.).

Although the complexity of the traffic situations of the autonomous vehicles is far lower than the complexity of controlling a car to navigate through various types of road traffic conditions to reach a desired destination, a few recent accidents in these simpler application domains provide valuable lessons. The common theme of the lessons we could learn from these examples is the mismatch between the human driver and the autopilot system in terms of their information processing and controlling abilities that is centred around the control of speed. More specifically,

Fig. 10.1 *Left* The scene after the accident. *Right* Top view of the bend with the crash cite indicated

the human driver either misunderstands the self-control mechanism of the vehicle or does not understand it at all. On the other hand, the automatic control system is not sufficiently intelligent to understand errors in parameter readings or mistakes made by the human controller. Also, the autopilot, the self-controller does not have the license to act even if the need for action is obvious because the human controller is the one who takes ultimate responsibility for the control. This seemingly obvious subordinate role for the autopilot may not always be the best control strategy in mixed control systems including self-driving cars, which various companies are developing.

10.2 Three Recent Accidents

10.2.1 Case 1: The Santiago de Compostela Derailment

The first example is a recent fatal train accident in Spain (Fig. 10.1). The rail disaster occurred on 24 July 2013, when an Alvia high-speed train was travelling from Madrid to Ferrol. This incident reveals major limitations of mixed control, which is often associated with automatic vehicle control systems. Mixed control system implies the presence of a driver who could take over the control of the vehicle by turning off the autopilot. The derailment of the fast train in a bend was due to the high speed, which was more than twice the speed allowed for entering that bend where the accident occurred. This human error could have been prevented or corrected with a simple automatic control mechanism. The automatic control system warned the driver three times to reduce speed but the driver did not react to these warnings and nobody, including the driver, knows why the warnings were ignored. The speed limit associated with the curvature of bends (or the strength and stability of the track) is the crucial parameter that played a central role in this tragic accident. The ability of humans to accurately perceive speed is fairly limited and perceived speed is known to be highly context dependent including speed adaptation (Owen and Warren 1987). Given the relatively slow rate of change and the long distance needed to reduce the speed of a high-speed train, a driver could easily make a mistake and the autopilot should be given a stronger control role rather than be subordinated under

these circumstances. Thus, the driver was not well calibrated to the control system used on the track and the automatic controller should have been better designed to allow fast acting correction of the speed when it was way outside the limit of safe speed range. In the immediate aftermath of the accident, three Automatic Braking and Announcement of Signals balises were installed on 1.9 km of the approach to Santiago de Compostela to enforce speed limits to prevent trains from reaching the 2013 accident point at a speed that would cause a similar derailment. Balises are track-mounted programmable transponders which communicate with the on-board computers on Spanish high-speed trains, and which can cause an automatic brake application if speed restrictions are not obeyed. These immediate actions indicate an indirect acknowledgement of a major design fault of the control system on this track.

10.2.2 Case 2: Air France Flight 447 Crash

Air France flight 447 from Rio de Janeiro to Paris crashed into the Atlantic in June 2009. A detailed analysis of flight recorders has revealed that airspeed sensors had malfunctioned (most likely because they had frozen up while flying through a thunderstorm). Again, the problem with the speed parameter (this time it was its measurement) triggered a chain of events that led to the loss of control of the aircraft. The captain was resting and perhaps the two less experienced co-pilots in control did not appear alert enough to deal with a high-altitude engine-stall. They ignored normal procedures and raised, rather than lowered, the plane's nose when it lost lift and stalled. Airbus A330 rolled from side to side in a tropical storm, and during a three and a half-minute plunge the pilots were unable to bring the aircraft under control. Even the captain who returned from a rest break was unable to regain control of the plane because the copilots were panicking and unable to tell him (and perhaps did not even know) what the problem was. Post-analysis suggests that the problem was again the lack of proper calibration of the co-pilots to deal with the speed control under the condition of malfunctioning speed sensors in a stormy weather. It was leaked through the media that all pilots were struggling with fatigue due to lack of enough sleep previous night and this might have contributed to the failure of dealing with the unusual conditions.

10.2.3 Case 3: Asiana Flight 214 Crash

On July 6, 2013, Asiana flight 214 in San Francisco struck a sea wall and broke apart on the runway following a missed approach. There were three fatalities and more than 180 other injuries. At the time of the accident, the airport navigation system was out of order because the Federal Aviation Administration was making runway safety improvements but the weather was clear and visibility was good. The captain was highly experienced in a Boeing 747 but was transitioning to flying a 777 and

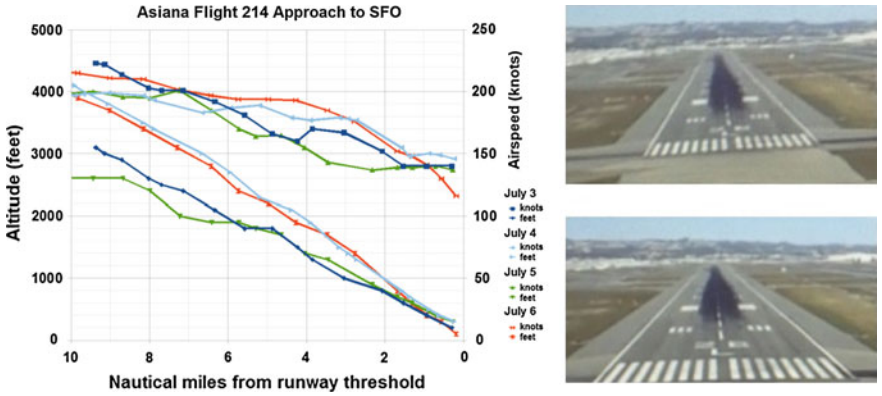


Fig. 10.2 The problem of landing a plane: The *left panel* shows the altitude and speed data of the last four landings of Asiana 214 (http://en.wikipedia.org/wiki/Asiana_Airlines_Flight_214) including the fatal one in San Francisco. The *right panel* shows views of two phases of landing on the same (L28) runway during a normal landing. These images imply optic flow with a *center* in the *middle* of images showing the expected landing area regardless of tilt that is indicative of side wind during the approach

admitted after the accident that he had been "very concerned" about landing and found it "very stressful, very difficult" (i.e., there were ethical concerns in using a mixed control system) to land without the glide-slope indicator that helps pilots determine whether the plane is too high or too low during approach. Although landing an aircraft is considered a dangerous operation that requires proper training, airline pilots are expected to be able to land an aircraft manually based on the visual control of approaching the landing area. Perhaps pilots tend to rely mostly on information (e.g., altitude, glide path, etc.) that are not salient for visual detection and they need more training for developing basic skills of using salient visual information (e.g., center and rate of optic expansion, Fig. 10.2) available in the optic flow (Gibson 1950, 1979).

Figure 10.2 presents the altitude and speed data of several landings of the same aircraft during July 3–6, 2013. The high variability of speed of approach is evident and despite similarities across landings, the fatal landing had a problem with speed control at the end phase at about one and a half nautical mile before landing. The speed should have been kept at about 140 knots, but it kept dropping down below 120. There was an attempt to abort landing and speed was increased but the speed problem was detected too late, just 7 seconds before the crash. Similarly to the Air France flight disaster, the speed control was involved in causing the accident in conjunction with the misunderstanding of the automatic system and the lack of sufficient experience. According to early reports, the safety board investigation was focusing on whether pilots have become overly reliant on automation to fly commercial planes, and whether basic manual flying skills have eroded. Investigators have also focused on the pilots understanding of the plane's auto-thrust system, which controls aircraft power. Some of the crew told investigators the auto thrust was always engaged, but

the safety board has said that system was not engaged. In general, again, there was a failure of the mixed control method with tragic consequences.

10.3 Lessons from These Examples

Although the control of trains and aircrafts are much simpler due to the simplicity of the environment relative to the complexity encountered during automobile driving in normal road traffic conditions, valuable lessons could be learned from the common aspects of these incidents. First of all, in all three cases, there was a mixed (human and automatic) control method implemented which resulted in a collective failure. Second, speed control can be regarded as a parameter calibration problem because the speed of motion has to be selected by taking into consideration the specific circumstances (e.g., curvature of bend being negotiated, mass of the vehicle, etc.). Third, the speed is linked to other parameters such as curvature/incline of bend, wind speed, altitude of flight and mass of the vehicle. These multiple parameters make the seemingly simple task of speed control quite complex for both artificial and human controllers. The control of a car in normal traffic is far more complex, because additional parameters such as road surface friction, angle of slope of the road surface, debris on the road, other traffic complexities, etc. should also be taken into consideration. Fourth, a long training and learning process is required in developing the ability to calibrate the parameters of vehicle motion to meet the actual environmental conditions for safe driving. Humans somehow manage to develop the skill of fast calibration of control parameters but they do not seem to rely on physical parameters such as speed, distance, resistance (airflow) etc. Finally, in order to develop smart artificial control strategies for automated vehicle control, it is advised that the control parameters are the same or similar to those used by human controllers and they should also be used in a similar way.

10.4 Parameters and Invariants of Human Visual Control

During World War II Gibson worked on problems of pilot selection, testing and training. He realized that the psychological theories of space perception (distance, speed, etc.) are useless and discovered what he called ‘optic flow patterns’ (Gibson 1950, 1979). For instance, in landing, while approaching a landing strip the surface point towards which the pilot is moving appears motionless (invariant) while all other surface points of the rest of the visual surface layout apparently moving away from that motionless surface point (focus of optic expansion) (see Fig. 10.2). Thus, the direction of motion could be controlled based on linking this invariant point and the intended point of touch down. Later, Lee (1976) has shown that drivers control speed in braking based on the rate of optic expansion. This control strategy is also linked to

an invariant, the so-called "tau-dot" which is close to 0.5 if the contact is to remain soft or the vehicle is to be stopped just before the contact surface.

The Asiana accident clearly suggests that pilots were not properly trained to use optic flow in visually controlled landing. Pilots were referring to the lack of glide path indicators from the traffic controller suggesting that they were trained to use artificial control parameters such as altitude, airspeed, etc. But these are cognitively complex and unnatural to process effectively in order to detect mistakes, which need correcting. Pilots should have been able to control landings based on a more natural strategy using optic flow and proper training for being calibrated to the parameters of Boeing 777 aircraft such as weight, engine power, etc.

In addition to optic flow-based invariants, several studies have demonstrated that behaviour control in various tasks is linked to other perceptual invariants the so-called ecological π-numbers scaled to body geometry (i.e., geometric measure of the human body or vehicle size) (Warren and Whang 1987). Shaw et al. (1995) proposed a generalization of this geometric scaling to invariant functions to accommodate various physiological and dynamic parameters (Invariant-Function Hypothesis. This is in sharp contrast with traditional speculative theories of perception, which tend to rely on physical parameters (e.g., distance, direction) of space and motion (e.g., speed) in behaviour control. Unfortunately, modern control theories seem to ignore Gibsonian insights and tend to rely on physical parameters (size, distance, speed, etc.) of traditional theories of perception that are shown to be inadequate to explain human perception (Owen and Warren 1987). These artificial control strategies could be learned but they are unnatural and too complex and under more demanding conditions such as the three examples presented (i.e., sensor error or missing glide-slope indicator) they could lead to major disasters.

10.5 Complexity of Driving

Car driving also requires the use of various perceptual invariants including optic flow-based ones as well as passability-related invariant. Specifically, the simple task of passing through an aperture (such as a gate) or between two other vehicles is a common task that could be used to demonstrate what strategies and perceptual invariants (π-numbers) humans use when artificial control with accurate sensors could develop seemingly much simpler and more straightforward control strategies.

10.5.1 Experiment on Linking Speed Control to Perceptual Invariants

Seven participants (age range: 21–26) drove a remote-controlled model car to perform a task of passing through a gate. Visual control was based on a monitor with the

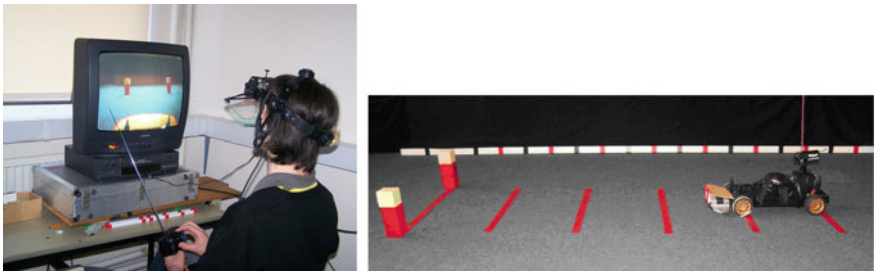


Fig. 10.3 *Left* A participant sitting in front of the screen, through which the car was visually controlled. Gaze recording was based on an ASL 501 eye-tracking system. *Right* The side-view of a trial as the remote-controlled car approached the gate

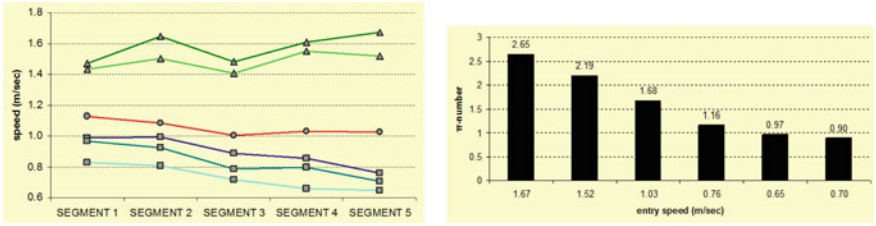


Fig. 10.4 Average performance for the six gates. *Left* The average speed for the 5 segments before the gate. *Right* The six associated π-numbers with the entry speed levels for the six gates

drivers' view (see Fig. 10.3). Six different gate widths (41, 34, 26, 18, 15 and 14 cm) were used. Four of these were wide enough to drive through, and the rest were narrower than the width of the car. The present experiment consisted of 24 randomized trials with each gate size presented four times. Participants were instructed to drive through accurately at the highest possible speed. Driving speed was examined as a function of gate-width. Five average speed values were measured in the last 100 cm (i.e., in five 20 cm-long segments) (Fig. 10.3). Entry-speed varied with gate size suggesting that different passability π-numbers (i.e., Aperture-Width/Car-Width ratio) are associated with different task dynamics (i.e., approaching and passing through speeds).

As expected, driving through gates of different width was associated with different speed profiles (Fig. 10.4). This finding confirms the invariant-function hypothesis for this task (Shaw et al. 1995). It is important to note, that the actual perceptual control processes are likely to be more complex and there is not enough research on this specific problem. We are currently investigating the problem of passability more closely (Kadar and Török 2012) but preliminary data suggest that there are individual differences, which might also be important if autonomous driving systems are to be

calibrated to the style of individual users. This is in line with Hodges (2007) approach that supports the need for a more articulated view of human behaviour based on values that could be sensitive to gender and cultural differences as well.

10.6 Conclusions

The present study has highlighted the differences between engineering control strategies, which are normally based on physical (kinetic and kinematic) parameters and human perceptual control strategies, which are usually based on perceptual invariants. This discrepancy is a potential source of errors with serious safety and ethical concerns, which has been demonstrated in train and flight control accidents. In developing autonomous car driving systems, the safety concerns and ethical consequences of these differences are expected to be more dramatic because of the complexity of traffic situations drivers encounter. It is especially important to highlight that artificial control mechanisms are focusing on calibration of parameters (dimensional numbers with distance, mass, etc., that are difficult to perceive for humans) while human perceptual control techniques mostly rely on perceptual invariants (dimensionless parameters, the so-called invariants or π-numbers), which are marginalized in theories of control mechanisms. Thus, the higher complexity of car driving tasks is more likely to enhance the difference between these two different strategies and making artificial car-control systems strange for human drivers. Obviously, these differences and our insufficient understanding of human driver performance would more likely lead to accidents than their use in much simpler control systems of trains and aircrafts. Given the fact that flight control training seems to overlook the natural invariant-based strategies, it is quite likely artificial car-control methods being developed are also full of artificial unnatural solutions that would make everyday human drivers nervous or frightened raising ethics concerns when mixed scenarios involve human drivers and autonomous cars in normal traffic situations. Long training procedures for getting used (calibrated) to these artificial car-control systems might help but cannot eliminate the danger of possible accidents and potentially unethical behaviours being exhibited (causing unnecessary stress, frustration, etc.). A paradigm shift is needed in control engineering in imitating human control strategies to minimize safety problems and ethical concerns in mixed control situations like we already have in train- and flight-control systems. Furthermore, it seems necessary to develop special displays (interfaces, Rasmussen and Vicente 1989) to facilitate visual control (and dynamic calibration) under limited visibility conditions such as driving at night, in heavy rain or dense fog. In general, the three case studies also suggest that mixed control could be improved by giving more power to artificial control to interfere if safety is being compromised. But to be able to design and implement a robust, safe and ethically compliant mixed human-artificial control system, a far better understanding of human control strategies is needed than our current knowledge can provide.

References

Gibson J (1950) Perception of the visual world. Houghton Mifflin, Boston

Gibson J (1979) The ecological approach to visual perception. Houghton Mifflin, Boston

Hodges B (2007) Values define fields: the intentional dynamics of driving, carrying, leading, negotiating, and conversing. Ecol Psychol 19:153–178

Kadar E, Török G (2012) Dynamic invariants in walking through an aperture while holding a tray with two hands. i-Perception 3. http://i-perception.perceptionweb.com/journal/I/volume/3/article/if598

Lee D (1976) A theory of visual control of braking based on information about time-to-collision. Perception 5:437–459

Owen D, Warren R (1987) Perception and control of self-motion: Implications for visual information of vehicular locomotion. In: Mark L, Warm JS, Huston RL (eds) Ergonomics and human factors: recent research. Springer, New York, pp 40–70

Rasmussen J, Vicente K (1989) Coping with human errors through system design: implications for ecological interface design. Int J Man-Mach Stud 31:517–534

Shaw R, Flascher O, Kadar E (1995) Dimensionless invariants for intentional systems: measuring the fit of vehicular activities to environmental lay-out. In: Flach J, Hancock P, Caird J, Vicente K (eds) Global perspectives on the ecology of human-machine systems. Lawrence Erlbaum Associates, Hillsdale, NJ, pp 293–357

Warren W, Whang S (1987) Visual guidance of walking through apertures: body-scaled Information for affordances. J Exp Psychol Hum Percept Perform 13(3):371–383

Chapter 11
Leader-Follower Strategies for Robot-Human Collaboration

L. Beton, P. Hughes, S. Barker, M. Pilling, L. Fuente and N.T. Crook

Abstract This paper considers the impact that robot collaboration strategies have on their human collaborators. In particular, we are interested in how robot leader/follower strategies affect perceived safety and perceived intelligence, which, we argue, are essential for establishing trust and enabling true collaboration between human and robot. We propose an experiment which will enable us to evaluate the impact of leader/follower collaboration strategies on perceived safety and intelligence.

Keywords Human-robot collaboration · Leader/follower strategies · Trust

11.1 Introduction

Human-robot collaboration has been on the research agenda for the robotics community for quite some time, but progress towards achieving it has been relatively slow. In the industrial setting it is now possible to have humans present within the workspace of a robot. This is due largely to the introduction of technologies such as force sensing which have enabled the introduction of safety features that prevent the human from physical harm. These robots will move out of the way of the human or stop when the human reaches into the workspace. This is still some way from true col-

L. Beton · P. Hughes · S. Barker · M. Pilling · L. Fuente · N.T. Crook (✉)
Oxford Brookes University, Oxford, UK
e-mail: ncrook@brookes.ac.uk

P. Hughes
e-mail: p.hughes@brookes.ac.uk

S. Barker
e-mail: stevebarker@brookes.ac.uk

M. Pilling
e-mail: mpilling@brookes.ac.uk

M.I. Aldinhas Ferreira et al. (eds.), *A World with Robots*,
Intelligent Systems, Control and Automation: Science and Engineering 84,
DOI 10.1007/978-3-319-46667-5_11

laboration. However, physical safety is not the only consideration when attempting to develop robots that are truly able to collaborate with humans. The establishment of trust lies at the heart of any such collaboration. In this work, we argue that trust in a robot depends at least in part on perceived safety and perceived intelligence, and that these, in turn, depend on the collaborative strategies that the robot adopts.

A significant number of studies have been performed on human-robot collaboration strategies. One of the key areas of interest is in the adoption of leader/follower (sometimes described as master/slave) roles within the collaboration. In these roles, it is only the follower/slave who adapts their behaviour to the leader/master (Kosuge et al., 1994). We seek to evaluate the impact that these two modes of collaboration have on the human's perception of the robot, particularly in terms of its perceived safety and intelligence. The evaluation involves a between subjects experiment in which a human and the Baxter robot collaborate in solving a series of marble maze puzzles. The human and the robot are each given control of one independent axis of rotation of the maze. Their task is to jointly guide a marble through the maze to one of a number of possible target positions. The robot is controlled using a Wizard of Oz (WoZ) technique. The evaluation was carried out using a post-experiment questionnaire that measures five key concepts of human-robot interaction including perceived safety and intelligence.

11.2 Related Work

A common perception of the collaborative relationship between humans and robots that is often portrayed in the media and popular fiction is of robots as the mechanical equivalent of slaves or servants of humans. However, a significant body of work has been published over recent years that examines this collaborative relationship in detail. Petersen (2007), for example, notes the difference between the slave and servant perception: slaves are generally forced to perform a task, whereas servants are generally willing to perform tasks as part of their act of service. He highlights the importance for robots to be designed with this goal in mind: they need to inherently 'want' to help humans. Throughout his paper Petersen compares the servitude of engineered robots with that of humans and concludes that this is not a fair comparison.

Bryson (2010), on the other hand, insists on breaking this analogy. In his view, robots cannot and should not be compared to humans, regarding them as nothing more than tools that provide extensions to the limitations of human capabilities. Even though they might display intelligence and apparent consciousness, both Petersen and Bryson argue that robots should not be regarded as possessing 'souls'. Bryson strongly recommends that no robot should be designed to appear to posses a soul or exhibit feelings. This view of robots, therefore, puts limits on how collaborative relationships can be described between them and humans; It seems unlikely that a human would regard themselves as collaborating with a tool to achieve a task.

Both Petersen and Bryson also comment on the issue of moral responsibility and conclude that a robot should not be treated as a moral agent, as it is unable to take

responsibility for any act. Although robots may appear at times to be autonomous, their apparent decisions are based solely on preprogrammed sequences. Only the operator or the manufacturer should be held responsible for the robot's actions. As such, a robot does not make ethical decisions autonomously. Rather, it should be designed to operate within predefined ethical bounds, to minimise potential harm. If accepted, this view of robots as lacking moral agency puts a limiting factor on how collaborative relationships with humans could be regarded; can a moral agent truly be described as collaborating in a task if it is unable to autonomously evaluate the moral impact of its contribution on its collaborators and on the successful completion of the task?

A recent study of the potential moral agency of robots is presented in Winfield et al. (2014). In their experiment, they program a robot with an internal model of itself, in order to simulate the consequences of the robot's next action. The robot is given a set of safety ratings, ranging from 'safe' to 'fatal', that are matched to outcomes of different ethical choices. The experimental setup has one robot programmed with the goal of protecting itself and one or two humans (simulated by robots) from falling into a hole. In the experiment with one simulated human, the robot demonstrated that it was capable of protecting itself and effectively move towards the human to prevent them from falling into the hole. In the experiment with two simulated humans, however, the robot is forced into a situation where it is likely that it will only be able to save itself and one of the humans from falling into the hole. On rare occasions, the robot by chance was able to prevent both humans from falling into the hole. The robots were not equipped with the capacity to decide "who should die", but their behaviour exhibited a minimal kind of moral agency which emerged out of the robots ability to predict the consequences of its actions.

Riek and Howard (2014) have proposed guidelines for robot design on a wider range of issues. They regroup recommendations into four main categories:

- Human dignity, or the respect of human rights and feelings;
- Design, considerations for optimal safety in both hardware and software;
- Legal, implying respect of current laws and regulations, as well as human informed consent;
- Social, to avoid deceiving or offending humans.

They note the tendency of humans to bond with other anthropomorphic and intelligent beings, and insist measures should be taken so robots behave ethically. Deception and emotional attachment play important parts in those guidelines.

In that context, it seems important to devise strategies to enforce such ethical rules. Kruijff and Janicek (2011) have proposed the use of doctrines to manage collaboration in mixed human-robot teams. They define doctrines as sets of abilities available for a given role, defining how different actors in a team should behave, in order to be both cohesive and ethical. Joint activity is ensured with the sharing of situational awareness: as they share common grounds and information, the team members are able to work together. The authors state that applying doctrines to each team member will permit mutual trust between agents, as well as entrusting a specific task to one agent. Since agents are given a responsibility through interaction with

others, they can thus be held responsible for the outcome of the task. The need for accountability for actions, as well as traceability is considered important, and was also mentioned by Riek and Howard (2014).

Jafar et al. (2014) have researched the emotional impact that collaboration with robots could have on humans. In three experiments, they ask humans to rate their feelings when in the presence of a moving robot. The first two experiments have humans simply stand next to the robot while it is operating on its own. The third one has the human subject program the robot behaviour, and then watch it working. It appears that humans tend to be cautious while a robot operates near them, but do not feel anxious. When the robot is 'collaborating', which is in that context executing orders, humans are still cautious but tend to feel impressed with the results.

In human-robot collaboration, the leader-follower paradigm has generated a great deal of interest from researchers. Some of this work focusses on how roles can be switched between the partners of a collaboration. The bulk of the research in that domain is based on physical manipulation. For example, Thobbi et al. (2011) developed a system for collaborating with a Nao robot to lift a table and keeping it level whilst it is carried. The robot is able to first follow the human's lead and simply reacts to the human's movements, but it is also able to predict the next movement of the human, and to take the lead by acting proactively. It switches between both modes depending on a confidence index which is increased when predictions match observations.

Whitsell and Artemiadis (2015) expand further on that approach, with a two-fold task. The goal is to keep a ball from rolling off a fixture, while moving that fixture to put the ball into the correct bin. To achieve this successfully, the tasks need to be performed simultaneously. Either the robot or the human can take the lead on either tasks. The robot is able to detect human intent by detecting changes in force. A role planner module decides which task the robot should take the lead on at any given time. This module makes decisions based on force thresholds, either fixed or adaptive. The researchers tested their system with three different thresholds, and had the robot programmed to move the ball to an incorrect bin half of the time, so the human could not stay passive if they wanted the task to succeed. The results were more significant when the threshold was updated throughout the task. Without adaptive thresholds, authors report that some subjects did not seem to perceive changes in leadership and tended to assume they were always leaders.

Other experimental work on collaboration has been done by Buondonno et al. (2015). In order to have a robot and a human dance a waltz together, they devised a predictive model system that is able to reliably predict the human's next move, in order to move the robot accordingly. The prediction here is based on either velocity or acceleration, with velocity showing best results. However, the robot is not yet able to switch roles and to lead the waltz by itself.

Leader-follower role switching was also experimented in other realms of collaboration. Weinberg et al. (2009) have proposed a way to allow a music band made of human and robot members to jam together. In follower mode, a robotic drummer and a robotic marimba player can detect beats and melodies from human players, and are able to 'follow' and play along. In leader mode they can improvise over

the detected melody. They switch between modes depending on the loudness and quantity of notes played by the human players. They reproduce a natural jam session further by making the robot's head nod along to the rhythm, and turning to different players to display interest and assimilate new melodies as they go.

Finally, Jarrasse et al. (2013) have reviewed the different approaches that have been taken to human-robot joint operation. Although their paper is named 'Slaves no longer', the views it expresses do not conflict with those expressed by Petersen and Bryson. While they do not question the assistive role of robots, they notice that the 'master-slave' paradigm is not the only one in current use and that more role options are available for collaboration. Common role distributions include not only leader-follower, but also partner-partner or independent team members, and teacher-learner. The latter is also a popular approach, facilitated by machine learning, where humans have to teach or train a robot's movement through physical or visual cues.

Jarrasse et al's survey shows that some researchers have compared human-robot interaction with the field of human-human interaction. They recognise collaboration with robots should make use of the same mechanics that are seen when humans collaborate with each other. However, this as yet has not been studied widely.

The authors observe that humans collaborate together by simulating an internal model of their partner, or their influence on the task at hand, much like robots do in the work of Winfield et al. However, if visual and haptic feedback is important for successful collaboration, they find that role distribution has more to do with the subject's nature or personality.

Thus, they argue that research should not only focus on motor factors, but also psychological and social ones. This seems to be crucial for understanding the nature of role switching. Authors want to highlight that role-switching might imply a certain level of equality between robots and human. Another hypothesis is that for specific tasks, competition might be more efficient than collaboration and should also be explored more.

11.3 Method

This study seeks to evaluate the impact that leader verses follower collaboration strategies have on the perceived safety and intelligence of robots. To enable this evaluation we developed a task in which both the human and the robot participant were equal partners in the contributions that they are able to make to the successful achievement of the task. The task should also enable one of the participants to take the lead and the other to follow their lead (or not). We have chosen the labyrinth game as the task for this evaluation. This game consists of a maze mounted on a horizontal board that can be independently tilted in a front-to-back ('pitch') and a side-to-side ('roll') direction using two control knobs. The aim of the game is to use the knobs to control the movement of a ball through the maze from a start to end position (Fig. 11.1). Normally the game is played by one person, who uses both knobs to control the pitch and roll movements of the board. In our experiments, one

Fig. 11.1 The Brio Labyrinth game

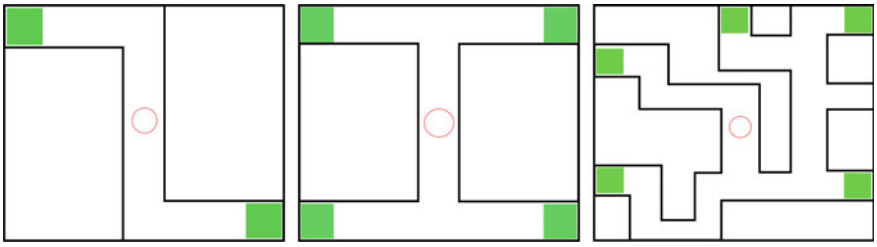


Fig. 11.2 Three maze designs ranging from simple to complex (*circles* indicate start positions, *green squares* indicate solutions

of the knobs was controlled by the human participant, and the other by the robot. Since the two knobs control independent movements of the board, the participants needed to collaborate to successfully take the ball through the maze from start to finish.

The mazes that we used for the experimental work ranged in difficulty from simple to complex (Fig. 11.2). Also, each maze had multiple solutions (i.e. multiple end points), which the human and the robot could navigate towards. This provides some uncertainty about which solution each participant in the collaboration is targeting. The robot was controlled by a remote operator using a 'Wizard of Oz' approach. There was no direct communication between the human and the robot or the operator.

In order to investigate the impact that leader and follower collaborative strategies have on perceived safety and intelligence we proposed the following experimental conditions:

A **The robot adopts a leader collaboration strategy throughout the game**.
In this condition, the robot operator will choose one of the available target solutions and control the robot to move the ball towards that solution. The human player, who may or may not choose to follow the robot's lead, is still able to influence the movement of the ball and hence the solution towards which it moves. The robot operator will need to take this into account as the game proceeds.

B **The robot adopts a follower collaboration strategy throughout the game**.
Here the robot operator observes the behaviour of the human player to try to identify which target solution the human player is attempting to reach. The operator will then control the robot to move the ball along the path leading to that solution.

C **The robot adopts a non-collaborative strategy throughout the game**.
In this condition the robot operator will seek to keep the ball at the start position on the board (i.e. away from any solutions). In other words, the robot will not collaborate with the human participant in moving the ball towards a solution. It is still possible in this condition for the human to successfully move the ball to a solution position in some of the mazes, but not with the active collaboration of the robot.

The participants are not informed that the robot is remote controlled by an unseen human operator until after they have completed the experiment.

11.3.1 Hypotheses

The hypotheses under investigation are that:

H1 The participants will rate the perceived safety of the robot in condition B more highly than condition A, and will rate condition A more highly than condition C.

H2 The participants will rate the perceived intelligence of the robot in condition A more highly than conditions B, and will rate condition B more highly than condition C.

11.3.2 Questionnaire

These hypotheses were evaluated through a post-experiment questionnaire. Several questionnaires exit in the literature that are designed to evaluate human perceptions of robots and significant work has been done to assure their reliability and validity (Bartneck et al. 2007; MacDorman 2006; Powers and Kiesler 2006; Ruijten and

Bouten 2014). In this study we used the Godspeed questionnaire to evaluate the perceived Intelligence and Perceived safety of the collaborative robot, as presented in Bartneck et al. (2008). The Godspeed questionnaire is made up of five sections that assess anthropomorphism, animacy, likeability perceived intelligence and perceived safety (Table 11.1). The 24 questions were randomised and the categories removed so as to disguise the fact that we are particularly interested in assessing perceived intelligence and perceived safety.

Table 11.1 The five "Godspeed" questionnaires

GODSPEED I: ANTHROPOMORPHISM						
Please rate your impression of the robot on these scales:						
Fake	1	2	3	4	5	Natural
Machinelike	1	2	3	4	5	Humanlike
Unconscious	1	2	3	4	5	Conscious
Artificial	1	2	3	4	5	Lifelike
Moving rigidly	1	2	3	4	5	Moving elegantly
GODSPEED II: ANIMACY						
Please rate your impression of the robot on these scales:						
Dead	1	2	3	4	5	Alive
Stagnant	1	2	3	4	5	Lively
Mechanical	1	2	3	4	5	Organic
Artificial	1	2	3	4	5	Lifelike
Inert	1	2	3	4	5	Interactive
Apathetic	1	2	3	4	5	Responsive
GODSPEED III: LIKEABILITY						
Please rate your impression of the robot on these scales:						
Dislike	1	2	3	4	5	Like
Unfriendly	1	2	3	4	5	Friendly
Unkind	1	2	3	4	5	Kind
Unpleasant	1	2	3	4	5	Pleasant
Awful	1	2	3	4	5	Nice
GODSPEED IV: PERCEIVED INTELLIGENCE						
Incompetent	1	2	3	4	5	Competent
Ignorant	1	2	3	4	5	Knowledgeable
Irresponsible	1	2	3	4	5	Responsible
Unintelligent	1	2	3	4	5	Intelligent
Foolish	1	2	3	4	5	Sensible
GODSPEEC V: PERCEIVED SAFETY						
Anxious	1	2	3	4	5	Relaxed
Agitated	1	2	3	4	5	Calm
Quiescent	1	2	3	4	5	Surprised

Fig. 11.3 The Baxter robot

11.3.3 Experimental Setup

We used a Baxter robot[1] for this study (Fig. 11.3). This robot is equipped with two 7 degree of freedom arms, each with an end effector and an RGB camera fitted at the end of the arm. One of the arms grips a control knob of the labyrinth game using an electric parallel gripper and is able to rotate the nob clockwise and counter clockwise. The other arm was positioned above the game board, with camera pointing downwards enabling the capture of live video of the maze as the game is being played (Fig. 11.4). This video is fed to the display of the robot operator so that they can see where the ball is in the maze and control the robot accordingly, depending on which of the three conditions is being followed.

The labyrinth game board has been modified from the commercially available Brio game. To capture the rotational information of the controller knobs, two linear potentiometers were physically linked to the transmission rod attached to the knobs (Fig. 11.5a). This provided X and Y rotation data. The potentiometers have a rotational range of 270°, and when fed with an input voltage, the output is a voltage

[1]Rethink Robotics: http://www.rethinkrobotics.com/baxter/.

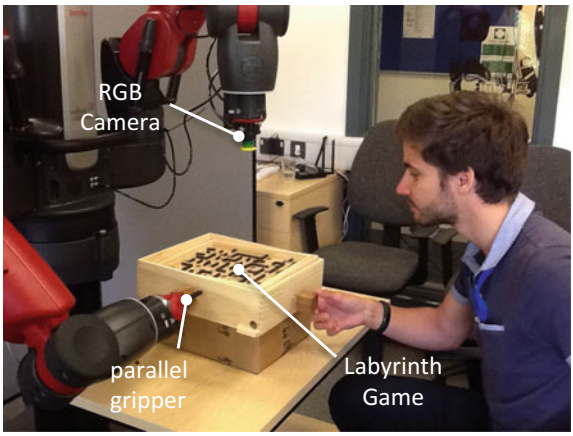


Fig. 11.4 The experimental setup

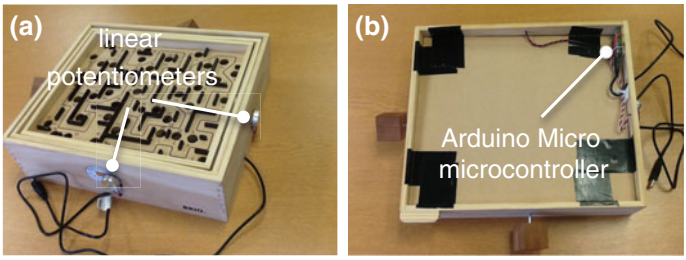


Fig. 11.5 The modified Labyrinth game board

bearing a linear relationship to the input voltage and position. Thus, if the potentiometer is fed 5 V and set to 135° (the halfway point), the output will be 2.5 V.

The voltages from the two potentiometers are fed into the A/D converters of the Arduino board, and the resultant digital data is returned to the system via a USB interface (Fig. fig:boardb). The system for providing positional data to the robot was developed using an 'Arduino Micro' microcontroller board, which is based on the ATmega32u4. This microcontroller is an 8 bit low power device running at 16 MHz with 32 KB of program memory. This board was used as it provides a 12-channel 10-bit A/D converter—returning integers in the range 0–1023, useful for converting the (original) analog positional information into a more useful digital form. Based on the fact that the A/D converter has a 10 resolution, and the maximum rotation of the potentiometers is 270°, this provides a system resolution of 0.26°, which is sufficiently accurate for the planned experiments.

11.3.4 Experimental Procedure

12 participants took part in the preliminary study. Each participant sat facing the robot with the labyrinth board between them (Fig. 11.4) and collaborated with the robot under one of the three conditions (A, B or C). Each participant played 15 games to give them the opportunity to become accustom to collaborating with the robot: 5 games with the simple maze layout (Fig. 11.4 top left), 5 games with the moderately complex maze layout (Fig. 11.4 top right), and 5 games with the complex maze layout (Fig. 11.4 bottom). Once a participant completed all 15 games, they filled out the post-experiment questionnaire and then they were debriefed. It is important to reiterate that participants were not informed about the remote control of the robot until after they had completed the questionnaire.

11.4 Ethical Considerations

Although this study of human-robot collaboration is based on a game, we believe that it has the potential to reveal some interesting insights into the impact that leader and follower collaboration strategies have on the perceived safety and intelligence of the robot. In this section of the paper we seek to make the case that (1) the perceived safety and perceived intelligence of robots are both fundamental for successful human-robot collaboration, and that (2) robots will need to adopt mixed leader-follower strategies to maximise their perceived safety and intelligence in collaboration tasks.

Both perceived safety and perceived intelligence of robots are essential in establishing the trust of human collaborators, and trust is important for successful collaborations. People will not trust a robot that, in their view, is not safe. Safety in the context of collaboration means not just personal physical safety, but also safety in terms of the collaborative task. A robot collaborating with a human to build a car, for example, might be viewed unsafe by the human if he/she has reason to think that the robot might damage the car during the collaboration.

Perceived intelligence is also essential for building trust in human-robot collaboration. In the context of a collaborative task, perceived intelligence relates to the robot's ability to act as a co-problem solver. That is, it relates to its ability to perceive a path to a solution, anticipate the next step in that path, and support the human in achieving that next step. If the human participant does not think that the robot possess this intelligence, then they are less likely to trust it as a collaborative partner. And so perceived safety and perceived intelligence are both necessary for establishing trust in collaborative tasks.

If it turns out that our proposed experiment confirms hypothesis H1, then robots who adopt the follower collaboration strategy are likely to be perceived as safer than those who follow the leader strategy. Robots who are leading in a collaboration may be perceived as presenting more of a threat than those who are simply following the human's lead.

If it also turns out that our proposed experiment confirms hypothesis H2, then this will support the view that robots that adopt leader collaboration strategy will be perceived as generally more intelligence than those who simply follow the human's lead. Robots who follow are not contributing to the problem-solving aspect of the task and are leaving this to the human.

If both H1 and H2 are supported by our experimental work, how, then, can a robot be perceived as being both intelligent and safe? We believe that this can only be achieved through a mixed leader and follower strategy, with the robot occasionally taking the lead at times, and following the lead of the human participant at other times. Precisely how the robot decides when to switch between leader and follower strategies is the subject of future work.

11.5 Results

The mean of the responses for each of the questionnaire items for each of the conditions are given in Fig. 11.6. This graph also includes an overall grand average of the participant ratings associated with intelligence (labelled 'overall intelligence') and safety (labelled 'overall safety'). It must be noted that a certain amount of caution needs to be exercised in interpretation of the results at this stage because the amount of participant data is insufficient to allow for formal statistical analysis. Nevertheless there are some clear indications of an emerging pattern between conditions for some of the ratings. In particular it can be seen that the robot when acting as a follower is viewed as being high in competence compared with when it is acting as a leader or as being obstructive; under these same conditions the robot is also viewed as being more knowledgeable than in the other conditions. Thus it seems that the follower strategy is one which in which the robot is viewed as most competent and knowledgeable about its given task. There was also a smaller trend for the robot to be judged as intelligent in the leader and follower collaboration strategies than in the obstruct strategy. The differences in the responses between conditions tended to be more apparent in some of the intelligence questions than the safety ones.

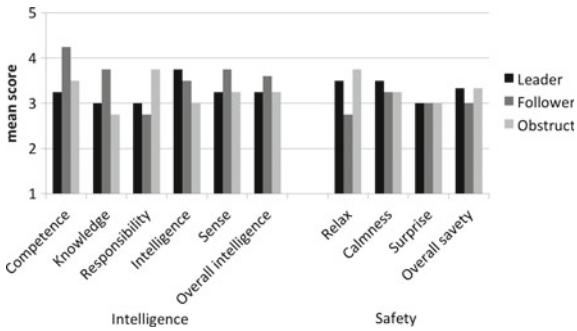


Fig. 11.6 Mean responses for each of the items in the leader, follower and obstruct conditions

In summary these preliminary result do indicate that the adopted collaboration strategy of the robot does have an influence the human operator's perceptions of the robot's character. The results seem to validate the importance of these aspects of a robot's behaviour in determining the quality of the relationship between a human operator and robotic collaborator on a manual task.

11.6 Concluding Thoughts

Human-robot collaboration continues to pose many challenges to the research community, both technical and ethical. In this paper, initial work on the study of leader and follower strategies for collaborating robots has been presented, showing in some detail the design of a proposed experiment to evaluate the impact that these strategies have on perceived safety and intelligence together with some very preliminary results. We have argued that perceived safety and intelligence are both fundamental to the successful human-robot collaboration and that a truly collaborative robot would employ a mixture of both strategies. Further experimental work is needed to properly evaluate this. Further work is also needed on how the robot decides on which strategy to adopt as the collaboration progresses.

References

Bartneck C, Croft E, Kulic D (2007) Is the uncanny valley an uncanny cliff? In: Proceedings of RO-MAN, interactive communication, Jeju, Korea, pp 368–373

Bartneck C, Croft E, Kulic D (2008) Measuring the anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of robots. In: Metrics for HRI workshop, vol 471

Bryson J (2010) Robots should be slaves. doi:10.2139/ssrn.1308500

Buondonno G, Patota F, Wang H, Luca A (2015) A model predictive control approach for the partner ballroom dance robot

Jafar F, Abdullah N, Muhammad M, Zakaria N, Ali Mokhtar M (2014) Investigation of human emotional state in human-robot collaboration. J Comput 9(3):668–677. doi:10.4304/jcp.9.3.668-677

Jarrasse N, Sanguineti V, Burdet E (2013) Slaves no longer: review on role assignment for human-robot joint motor action. Adapt Behav 22(1)

Kosuge K, Yoshida H, Taguchi D, Fukuda T, Hariki K, Kanitani K, Sakai M (1994) Robot-human collaboration for new robotic applications. In: Proceedings of IECON'94—20th annual conference of IEEE industrial electronics, vol 2. doi:10.1109/IECON.1994.397872

Kruijff G, Janicek M (2011) Using doctrines for human-robot collaboration to guide ethical behavior

MacDorman K (2006) Subjective ratings of robot video clips for human likeness, familiarity, and eeriness: An exploration of the uncanny valley. In: Proceedings of ICCS/CogSci-2006 long symposium: toward social mechanisms of android science. Vancouver

Petersen S (2007) The ethics of robot servitude. J Exp Theor Artif Intell 19:43–54. doi:10.1080/09528130601116139

Powers A, Kiesler S (2006) The advisor robot: tracing people's mental model from a robot's physical attributes. In: Proceedings of 1st ACM SIGCHI/SIGART conference on human-robot interaction, Salt Lake City, Utah, USA

Riek L, Howard D (2014) A code of ethics for the human-robot interaction profession. In: We robot conference

Ruijten P, Bouten D (2014) Introducing a rasch-type anthropomorphism scale. In: Proceedings of 2014 ACM/IEEE international conference on Human-robot interaction, pp 280–281

Thobbi A, Gu Y, Sheng W (2011) Using human motion estimation for human-robot cooperative manipulation. In: Proceedings of IEEE international conference on intelligent robots and systems, pp 2873–2878. doi:10.1109/IROS.2011.6048572

Weinberg G, Blosser B, Mallikarjuna T, Raman A (2009) The creation of a multi-human, multi-robot interactive Jam session. In: Proceedings of international conference on new interfaces for musical expression, pp 70–73

Whitsell B, Artemiadis P (2015) On the role duality and switching in human-robot cooperation : an adaptive approach

Winfield A, Blum C, Liu W (2014) Towards an ethical robot: internal models, consequences and ethical action selection. In: Advances in autonomous robotics systems. Lecture notes in computer science, vol 8717. Springer, pp 85–96

Chapter 12
Industrial Robot Ethics: The Challenges of Closer Human Collaboration in Future Manufacturing Systems

S.R. Fletcher and P. Webb

Abstract As a result of significant advances in information and communications technology the manufacturing industry is facing revolutionary changes whereby production processes will become increasingly digitised and interconnected cyber-physical systems. A key component of these new complex systems will be intelligent automation and human-robot collaboration. Industrial robots have traditionally been segregated from people in manufacturing systems because of the dangers posed by their operational speeds and heavy payloads. However, advances in technology mean that we will soon see large-scale robots being deployed to work more closely and collaboratively with people in monitored manufacturing sytems and widespread introduction of small-scale robots and assistive robotic devices. This will not only transform the way people are expected to work and interact with automation but will also involve much more data provision and capture for performance monitoring. This paper discusses the background to these developments and the anticipated ethical issues that we now face as people and robots become able to work collaboratively in industry.

Keywords Human-robot collaboration · Manufacturing · Industrial robots · Industrial automation

12.1 Introduction

The manufacturing industry has a long history of pursuing technological development and innovation for continuous improvement, but it is commonly accepted that have been three major periods of transformational change that most scholars consider as 'industrial revolutions'. The first industrial revolution (c.1760–1830) refers to the long period of progress in which the development of water, steam power and

S.R. Fletcher (✉) · P. Webb
Centre for Structures, Assembly and Intelligent Automation,
Cranfield University, Cranfield, UK
e-mail: s.fletcher@cranfield.ac.uk

M.I. Aldinhas Ferreira et al. (eds.), *A World with Robots*,
Intelligent Systems, Control and Automation: Science and Engineering 84,
DOI 10.1007/978-3-319-46667-5_12

machinery which led to the centralisation of production within factories (Ashton 1966). In the second revolution (c. 1870–1914) electrical power and the invention of the moving conveyor enabled assembly lines and mass production to create growth in economies of scale (Mokyr 1998). The third 'digital revolution' brought change to production methods via advances in electronics, computing and automation from the middle of the 20th century and some authors argue that this is still ongoing, because information technology and digitisation is still being developed (Rifkin 2016). However, it is now widely recognised that we are now entering a new and distinct phase, the fourth industrial revolution, in which all of these prior technologies are now being integrated in new and transformational ways (Schwab 2015). Labelled 'Industrie 4.0' by German national initiatives to strengthen manufacturing competitiveness (Hermann et al. 2015), this new industrial revolution will see factory production systems transformed by digitisation and increasingly integrated technologies.

Intelligent information systems and 'cloud manufacturing' are going to convert traditional sequential man-machine production processes into more sophisticated and digitally inter-connected cyber-physical systems for the optimisation of flexible manufacture and service (Kagermann et al. 2011). Emerging internet and informatics technologies will not only communicate with each other internally but also with other systems and organisations external to the factory (Jacobs 2015). The advancing sophistication of computer mediated communications and data-driven internet based networks will also provide adaptive control of production and service systems in real time (Yin and Kaynak 2015) and intelligent interactions between physical and autonomous elements which will enable the greater application of robots on the shop floor (Wan et al. 2015; Lanza et al. 2015). Production systems will therefore involve much greater integration of humans and automation via various components such as informatics, robotics, mobile devices and sensors. Although it is claimed that "Industrie 4.0 is not initiated on a shop floor level" (Schuh et al. 2015), it seems inevitable that the changes brought about by this movement will transform shop-floor environments. This will have significant implications for the remaining human workforce.

Advancing technologies are relevant to almost every area of life, and humans will need to adapt to increasing levels of intelligent automation, informatics and cyber-physical systems in their daily protocols and encounters. However, the manufacturing context for human-robot relationships is distinct and is facing upheaval as a result of the dynamic Industrie 4.0 revolution. This chapter specifically addresses that context to discuss the ethical issues and requirements related to the augmentation of robotics and human collaborations. Specifically, in order to consider why and how human-robot collaboration is likely to bring ethical impacts and issues we address three key issues across the following sections:

- The current state of industrial human-robot collaboration
- The current state of human issues and ethics
- The likely impacts to arise from future human-robot collaboration systems?

Together these questions will reveal the reasons why robot ethics need to be considered for the industrial context. Finally, we conclude this chapter by drawing

together key points and making recommendations for both future research and for the optimisation of the design and operational success of future human-robot systems in manufacturing system.

12.2 The Current State: Industrial Human-Robot Collaboration

Increasing the use of robots in production systems has long been a goal of manufacturing organisations. Industry has consistently viewed people as unreliable, unpredictable and expensive since productivity benefits from deskilling and simplifying human work via the division of labour became apparent in the early days of mass production (Mintzberg 1983). This led to a pervasive "machine school of thought" in industry where machinery was prioritised over people and organisations pursued the development of automated processes to replace human work wherever possible (Doyle 2003). However, despite long industrial aspirations for eliminating humans completely to create fully automated 'lights out' factories there have been very few opportunities to do so. Most products still require some sort of human input in their manufacture. Some tasks require physical dexterity and agility but many others require the flexibility and intellectual reasoning which a robot is not yet able to supply (Shen et al. 2015; Ding and Hon 2013). So, although the manufacturing industry has focused on automating systems it will continue to require human input.

To date, robots have generally been employed in manufacturing for heavier, more repetitive work: "to replace human workers performing dangerous, difficult, dull, monotonous and dirty tasks" (Vasic and Billard 2013). This means industrial robots are usually of medium or large payload and have had to be segregated from the workforce on the shop floor to protect people from the potential hazards posed by their considerable strength and speed. Consequently, in manufacturing system layouts they have typically been positioned upstream to exploit their superior strength and reliability for the more simple and unskilled operations, typically at fully automated stations (Hedelind and Kock 2011), whilst human operators are positioned to apply dexterous and cognitive skills on more complex and skilled assembly tasks further downstream (De Krüger et al. 2009). To comply with the segregation requirements of health and safety regulations and standards, these robots have also been set apart from people by the use of various forms of physical guarding, such as fencing/'caging', sensor based safety light curtains, and other separation distance measures. As this has been the convention in industry for such a long time workforces are highly accustomed to perceiving robots as hazardous.

Although physical separation of robots in manufacturing layouts has been a relatively straightforward measure to protect people from the danger of collision it is not ideal for reasons of safety and productivity. First, physical boundaries are not failsafe boundaries and most accidents and injuries occur when a human has entered the work cell zone while the robot is operating (Vasic and Billard 2013). Second,

separation prevents direct interaction and collaboration between the human and the robot (Kulić and Croft 2006) and can be a cause of considerable costly disruptions to work flow efficiency. The need for physical segregation means robots are typically only useful in areas of the factory where there is little or no human input and it prevents robots from being positioned in many areas near to the workforce where they could be of benefit and assistance to human operators. Thus, it would be preferable if humans and robots were enabled to work more closely together at points in production systems where skilled and non-skilled tasks are required. This concept of industrial human-robot collaboration refers to the desirable real time union of human and robot in a shared workspace such that production activities are continuous and uninterrupted.

The concept of combining humans and robots to utilise the attributes of each in joint tasks is not new, but is becoming increasingly possible as advances in the sophistication of control algorithms, metrology, sensory capability and actuators make it possible to produce integrated safety systems which can monitor a designated workspace (De Santis and Siciliano 2008). Ongoing research has demonstrated the potential for integrating such technologies to create vision-based systems with all-round (360° and topographical) ubiquitous monitoring capabilities (Walton 2013). As these systems would activate theoretically should be safer than traditional physical guarding will allow operators to work in closer proximity with medium or large payload robots safely. Additionally, there is also now a greater availability of force- and torque-limited robots which are deemed safe and suitable for close and direct interaction with humans as their collision impact is mild enough not to cause 'unacceptable' harm according to current guidelines and understanding (HSE 2013). This will allow operators to work with smaller and lighter payload robots and enable the implementation of robotic assistants in a wider range of potential locations and applications than ever before.

12.3 The Current State: Human Issues and Ethics

As discussed, there has been a longstanding imbalance in the manufacturing industry where the advancement of technology has been prioritised over efforts to understand the impact it will have on the people who will be expected to interact with it. Evidence from various real industrial cases of advanced manufacturing technology implementation have revealed that a lack of consideration of human/user issues has been a root cause of failure in a number of unsuccessful cases (Chung 1996). Research has also showed that a lack of attention to human issues and interactions in the design/redesign of automated systems can impact on overall production performance (Fletcher et al. 2008). This sort of evidence has led to some improvement in understanding the importance of integrating human factors into the engineering design of industrial systems and work processes (Battini et al. 2011) but systems are still generally designed with a lack of balanced sociotechnical system design approaches (Baxter and Sommerville 2011). It has been noted that despite that industrial engineering has been gathering a

"solid science base" of human factors knowledge in relation to automation over many years (Parasuraman and Wickens 2008) the "engineers who develop robots" still tend to lack understanding and/or application of relevant psychological principles such as cognition and perception (Bartneck et al. 2008).

Increasing automation in modern production systems has "already been taken far without paying sufficient attention to the specific knowledge, skills and abilities of the human operator" (Mayer et al. 2011). As it has been shown that new automation and technology in production systems does not completely replace human work but, instead, changes it to satisfy the needs for supervision and maintenance roles (The Economist 2012) there will inevitably be a need to also consider the implications of such changes. Recent work has begun to identify the key factors likely to influence organisational implementation of human-robot collaboration as well as the individual-level factors likely to affect workers' adoption of, and trust, in it (Charalambous et al. 2015a) but little is being done to address ethical issues.

Given the evidence and indications, it is reasonable to expect that attention to the human issues and ethics that will arise as a result of Industrie 4.0 human-robot systems will continue to lag behind the level of attention that will be paid to the design and implementation of the emerging technologies. It is often the case that engineering design approaches rely on compliance with safety standards and fundamental ergonomic design principles to cover their attention to human issue but that does not incorporate wider issues such as ethics. So, if we are to learn from the industrial mistakes of the past and ensure sufficient appreciation of human issues and impacts is incorporated into the design of new systems it is vital that we now start to raise awareness of ethical implications.

12.4 Specific Ethical Issues

A significant consequence of the impending Industrie 4.0 revolution will be the degree to which human operators will be required to adapt to a vast number of physical, psychological and social changes to their workplace environments and to the nature of the work they do. As robots and intelligent automation will play a critical role in 'cloud manufacturing' and cyber-physical systems (Givehchi et al. 2013) it is inevitable that organisations will continue trying to increase the number of robots they deploy. However, it is clear that human input will also need to be maintained and ideally employed to work collaboratively with the robots. To avoid the pitfalls of past technology adoption failures implications of these changes should be considered and designed for in advance, to support the wellbeing of the workforce and to optimise the chances of technological and operational success.

As the robotics in new production systems will be integrated with a range of other emerging technologies such as wireless sensor networks, big data, cloud computing, embedded systems, mobile internet, etc. (Wang et al. 2016) the impacts on workforces will be multifarious. In this chapter we are focusing primarily on the potential ethical

issues surrounding human-robot collaboration but, given the interconnectivity of future systems, there will inevitably be overlaps with other associated technologies.

As discussed, there it is now widely accepted that people will still be required to work with/in future manufacturing systems, in spite of increased levels of automation and robotics. Indeed, it is the very escalation of automation that will bring one of the biggest changes to the human role as operators will need to be retrained and upskilled to replace their manual production tasks with more supervisory and monitoring roles (The Economist 2012). With human-robot collaboration being a key feature of future production systems workers will be required to engage in symbiotic relationships with industrial robots (Wang et al. 2015). So, it is important that ethical issues are considered with respect to the management of this significant change so that we can help operators move from traditional manual methods to new supervisory and auxiliary roles with the minimum of disruption and negative affect, and to maximise worker acceptance and wellbeing. Although there are a great many potential issues to be deliberated and discussed in this chapter we consider three key areas: protection from harm, informed choice and confidentiality and performance data monitoring.

12.4.1 Protection from (Psychological) Harm

Firstly, the new expectation for a human to even venture near to an industrial robot will require a new outlook and level of trust. The traditional use of physical measures to segregate industrial robots from operators will have established distinct and visible proximity boundaries, sending out an abundantly clear message that the industrial robot is a danger and is to be avoided and kept at distance. This message will also have been reinforced via training, operational protocols and safety communications. Consequently, even though new ubiquitous monitoring systems can be inherently safer than traditional safeguarding, industrial workforces will to some extent already be psychologically primed to mistrust robots and expect protective segregation. This means they may be uncomfortable with direct interaction and proximity, and may even resist robot deployment on the shop floor. How easily will workers be able to adapt to new expectations for collaboration and become familiarised and trusting of the surrounding ubiquitous monitoring system? Research has begun to understand trust in industrial automation, and its psychological dimensions (Charalambous et al. 2015b), but we still need to understand more to ensure the concept is introduced carefully, as operators may need to adapt long-standing psychological attitudes towards robots.

The rising availability of smaller and lighter payload robots may already be initiating a change in attitudes to industrial robots. However, although their safety is supported by studies that have determined ‘acceptable’ levels of collision impact these focus on physical outcomes. In terms of psychological impacts there is little evidence to tell us how workers will feel about the constant possibility of experiencing ‘safe’ collisions. How does daily exposure to this low risk impact on an operator’s satisfaction and wellbeing, or on their performance? Similarly, as robots are being

developed with increasing levels of autonomy and mobility, and advanced capabilities for self-learning and adaptation, operators may not only have to familiarise with, and accept, robot autonomy/adaptivity but also adapt their own behaviour in return. How will operators adjust to working amongst autonomous robots and develop productive relationships with them? Once again, to ensure ethical standards that promote both worker acceptance and wellbeing we need to consider the implications of imposing frequent exposure to 'acceptable' levels of collision risk and close proximity to small/autonomous robots.

Ethically then, the dominant focus on protecting workers from physical harm must continue but we now also need to consider the new potential for psychological harm in intelligent systems. Simply asking individuals to change their assumptions of danger or accept prescribed levels of risk acceptability may infuse various psychological concerns and anxieties, and we should not assume they are able to adopt new concepts that directly conflict with long-held beliefs. Some attention to psychological impacts has been included in the new UK standard: 'Robots and robotic devices: Guide to the ethical design and application of robots and robotic systems' but this is not context-specific and therefore may not sufficiently cover industrial human-robot collaboration. In organisation, ethical considerations and measures for psychological protection need to be incorporated into training, support and remedial interventions—but to do this we need to first improve our understanding of impacts linked to appropriate adjustment and coping strategies.

12.4.2 Informed Choice and Deception

As outlined, we need to improve our understanding of the impacts and changes that will be to be brought about by new Industrie 4.0 human-robot systems. In order to foster a new outlook and sense of trust from operators it may be important to provide training with fairly comprehensive information about the functionality and reliability of human-robot collaboration systems so that they are not only fully aware of user protocols but of levels of risk and mitigation functions. Without enlightening operators, will there be a greater risk of lack of trust and acceptance? What is the best level and type of information to provide so that operators can grasp the key concepts of new systems and develop trust? Ethically, it would be ideal if users were made aware of all of the technical facts and what is going to be expected of them, without any form of misinformation or deception, so that they are able to make an informed choice whether to accept the new role change. Although there is an understandable limit to how much detail is relevant it would seem ethically appropriate– and probably conducive to worker acceptance—if a suitable level of information is provided.

Amongst the ethical issues concerning the introduction of such a new way of working, and the requirements for attitude change/re-education that it will bring, there is also a question of voluntary redeployment and role change. Although workers must obviously comply with reasonable practices and procedures laid down by the organisation it would not be advantageous to force anyone to work in a system

they do not trust or no longer trust. Such anxieties would be detrimental to worker wellbeing and performance. What would be the most suitable way to introduce and manage significant change in work practices and expectations of working collaboratively with robots? What is the best way to manage voluntary operator's skills and training requirements for such role changes and manage the deployment of reluctant or resistant workers? Little or no work has yet been undertaken to advise on these best methods of change management.

12.4.3 Confidentiality and Performance Data Monitoring

Whereas previous industrial revolutions have focused on reducing power and increasing productivity, the impending 'Industrie 4' revolution focuses on the dynamic flow of information (Essers and Vaneker 2015). Factories of the future will continuously collect, distribute and maintain 'big data' via various component systems including robotics. Robots designed for interactions with humans will no longer be designed simply for joint 'turn-taking' tasks but will involve more complex interactions with people and other system elements. It is inevitable that in many cases data will not just be delivered to a human-robot system but will also be collected from it, either deliberately as part of system performance monitoring or just as a by-product of the system's inter-connected informatics. This means that interfaces in more highly digitised systems are likely to capture human data directly or indirectly, overtly or covertly. Performance data monitoring is a sensitive topic so the way in which new industrial systems will collect and use such data from individuals is a critical issue.

Obviously any data which could be linked to identifiable individuals should still comply with data protection protocols but these would normally require subject awareness and permission. How will people be made aware of all data being collected about/from them and will they be required to provide (sufficiently) informed consent? To what extent will data on individual performance be gathered and appraised? It is not yet clear how such systemic data capture will differ from current performance data management practices, and how it may need to be managed differently to maintain personal protection and ethical suitability. This is another issue that needs to be considered and addressed appropriately to optimise workforce trust and acceptance, as people need to be confident that there is no inappropriate intrusion into their privacy or rights.

12.5 Conclusion and Recommendations

This chapter has briefly described a little of the historical backdrop to the application of industrial robots and the current state in which human-robot collaboration in manufacturing is a growing movement and human issues and ethics need to be addressed. It can be seen that there has been a long tradition of prioritising technology

over people and efforts to replace manual tasks with automation wherever possible, but now an acceptance that people will need to be retained and will be required to work collaboratively with intelligent robots in future systems. The ethical implications of these new expectations have not yet been sufficiently explored and understood. It has also been discussed that, as a result of safety measures, manufacturing workforces are accustomed to a prevailing concept of robots in manufacturing as hazardous and segregated from their workspace, but due to impending advances in human-robot collaboration workers will need to adapt their beliefs to accept and trust in system safety. Importantly, this chapter has also demonstrated that there has been a prevalent lack of attention to human issues which has contributed to various failures in technology acceptance, which means we need to better consider such issues and ethics as we enter the fourth industrial revolution. In answer to the current state a set of key ethical issues and associated recommendations are proposed.

First, we need to undertake robust studies to help us understand and minimise potential psychological impacts of new human-robot collaborations in industry. Specifically, we should explore how to enable operators to trust and understand the concept of collaborative robots and ubiquitous monitoring safety systems, how to manage operators' exposure to robots with 'acceptable' levels of risk and autonomy, and how to design training and support to cover these issues.

Second, we need to establish a suitable level and delivery of information to enable operators to make an informed choice to work with robots, and to develop their trust them. It is also important that organisations understand how to manage such significant changes and enable people to adapt to new roles voluntarily and satisfactorily.

Third, as potentially sensitive performance data is going to be collected as part of human integration in collaborative systems it is critical that this data is collected, stored and used ethically. Moreover, we need to determine the best method for ensuring personal data is protected so that privacy is upheld.

Overall, research is beginning to reveal some of the key factors and psychological impacts of industrial human-robot collaboration (Charalambous et al. 2015a, b) but the development of this body of knowledge is still in its infancy. Ethical issues need to be investigated and incorporated with emergent findings form human research to establish ethical guidelines for the design and implementation of human-robot collaboration systems. As the Industrie 4.0 movement advances formal ethical guidance is needed as early as possible (to avoid the pitfalls of neglecting human considerations). The issues presented here are not intended to be comprehensive but to raise awareness for future development and debate in the aim of encouraging safe and ethical design of industrial robot systems.

References

Ashton T (1966) The industrial revolution 1760–1830. In: Hands of a child, vol 109. Oxford Academic Press

Bartneck C, Croft E, Kulic D (2008) Measuring the anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of robots. In: Proceedings of metrics for human-robot interaction workshop in affiliation with the 3rd ACM/IEEE international conference on human-robot interaction (HRI 2008). Technical report 471. University of Hertfordshire, 37–44, Amsterdam

Battini D, Faccio M, Persona A, Sgarbossa F (2011) New methodological framework to improve productivity and ergonomics in assembly system design. Int J Ind Ergon 41(1):30–42

Baxter G, Sommerville I (2011) Socio-technical systems: from design methods to systems engineering. Interact Comput 23(1):4–17

Charalambous G, Fletcher S, Webb P (2015a) Identifying the key organisational human factors for introducing human-robot collaboration in industry: an exploratory study. Int J Adv Manuf Technol:1–13

Charalambous G, Fletcher S, Webb P (2015b) The development of a scale to evaluate trust in industrial human-robot collaboration. Int J Soc Robot:1–17

Chung C (1996) Human issues influencing the successful implementation of advanced manufacturing technology. J Eng Technol Manage 13(3):283–299

De Krüger J, Lien T, Verl A (2009) Cooperation of human and machines in assembly lines. CIRP Ann Manuf Technol 58:628–646

De Santis A, Siciliano B (2008) Safety issues for human-robot cooperation in manufacturing systems. In: Tools and perspectives in virtual manufacturing, Napoli, Italy, July

Ding Z, Hon B (2013) Constraints analysis and evaluation of manual assembly. CIRP Ann. Manuf Technol 62(1):1–4

Doyle C (2003) Work and organizational psychology: an introduction with attitude. Psychology Press, Hove

Essers M, Vaneker T (2015) Design of a decentralized modular architecture for flexible and extensible production systems. Mechatronics

Fletcher S, Baines T, Harrison D (2008) An investigation of production workers' performance variations and the potential impact of attitudes. Int J Adv Manuf Technol 35:1113–1123

Givehchi O, Trsek H, Jasperneite J (2013) Cloud computing for industrial automation systems—a comprehensive overview. In: Proceedings of 2013 IEEE 18th conference on emerging technologies & factory automation (ETFA), pp 1–4

Hedelind M, Kock S (2011) Requirements on flexible robot systems for small parts assembly, a case study. In: Proceeding of international symposium on assembly and manufacturing, 25–27 May, Tampere, Finland

Hermann M, Pentek T, Otto B (2015) Design principles for industrie 4.0 scenarios: a literature review

HSE (2013) Collision and injury criteria when working with collaborative robots (RR906). Health and safety executive 2013

Jacobs R (2015) Rise of robot factories leading 'fourth industrial revolution'. In: Newsweek (US edition), March 5, 2015, Newsweek. http://www.newsweek.com/2015/03/27/rise-robot-factories-leading-fourth-industrial-revolution-311497.html. Accessed 21 Jun 2015

Kagermann H, Lukas W, Wahlster W (2011) Industrie 4.0: Mit dem Internet der Dinge auf dem Weg zur 4. industriellen Revolution. VDI nachrichten 13

Kulić D, Croft E (2006) Real-time safety for human-robot interaction. Robot Auton Syst 54:1–12

Lanza G, Haefner B, Kraemer A (in press, 2015) Optimization of selective assembly and adaptive manufacturing by means of cyber-physical system based matching. CIRP Ann Manuf Technol

Mayer M, Schlick C, Ewert D, Behnen D, Kuz S, Odenthal B, Kausch B (2011) Automation of robotic assembly processes on the basis of an architecture of human cognition. Prod Eng 5(4):423–431

Mintzberg H (1983) Structuring by fives: designing effective organizations. Prentice-Hall, Englewood Cliffs, NJ

Mokyr J (1998) The second industrial revolution, 1870–1914. In: Valerio C (ed) Storia dell'economia Mondiale. Laterza Publishing

Parasuraman R, Wickens C (2008) Humans: still vital after all these years of automation. Hum Factors J Hum Factors Ergon Soc 50(3):511–520

Rifkin J (2016) The 2016 world economic forum misfires with its fourth industrial revolution theme. In: The economist, Jan 15, 2016, the economist. http://www.industryweek.com/information-technology/2016-world-economic-forum-misfires-its-fourth-industrial-revolution-theme. Accessed 20 Apr 2016

Schuh G, Reuter C, Hauptvogel A, Dölle C (2015) Hypotheses for a theory of production in the context of industrie 4.0. In: Advances in production technology. Springer, pp 11–23

Schwab K (2015) The fourth industrial revolution. Foreign Affairs 12

Shen Y, Reinhart G, Tseng M (2015) A design approach for incorporating task coordination for human-robot-coexistence within assembly systems. In: Proceedings of 9th IEEE annual IEEE international systems conference (SysCon), pp 426–431

The Economist (2012) Making the future: how robots and people team up to manufacture things in new ways. Economist. http://www.economist.com/node/21552897, April 21st, 2012. Accessed 30 May 2015

Vasic M, Billard A (2013) Safety issues in human-robot interactions. In: Proceedings of 2013 IEEE international conference on in robotics and automation (ICRA), pp 197–204

Walton M (2013) In the context of wing equipping—a framework for safeguarding direct cooperation between high load, industrial robots and human operators. Unpublished PhD thesis

Wan J, Cai H, Zhou K (2015) Industrie 4.0: enabling technologies. In: Proceedings of 2014 IEEE international conference on intelligent computing and internet of things (ICIT), pp 135–140

Wang L, Törngren M, Onori M (2015) Current status and advancement of cyber-physical systems in manufacturing. J Manuf Sys (Part 2) 37:517–527

Wang S, Wan J, Li D, Zhang C (2016) Implementing smart factory of industrie 4.0: an outlook. Int J Distrib Sens Netw

Yin S, Kaynak O (2015) Big data for modern industry: challenges and trends [point of view]. Proc IEEE 103(2):143–146

Chapter 13
Clarifying the Language of Lethal Autonomy in Military Robots

Sean Welsh

Abstract Many argue that robots should not make the decision to kill humans and thus call for a ban on "killer robots" or lethal autonomous weapons systems (LAWS). However lethal decision making is complex and requires detailed analysis to define what is to be banned or regulated. It is common to make distinctions between in the loop, on the loop and off the loop LAWS. It is also common to refer to the "critical functions" of selecting and engaging targets. In this paper I propose two extra LAWS types. A Type 0 LAWS is an RPV with "no robot on the lethal loop." A Type 4 LAWS is a robot that has gone "beyond human control" and has "no human in the loop." Types 1–3 are the familiar in, on and off the loop LAWS. I also define a third "critical function" namely defining the targeting criteria. The aim is to clarify what exactly is meant by "meaningful human control" of a LAWS and to facilitate wording such as might occur in a Protocol VI to be added to the Convention on Certain Conventional Weapons (CCW).

Keywords Robot ethics · Lethal autonomous weapons systems · Killer robots · Meaningful human control · International humanitarian law · Machine ethics

13.1 What Is a Killer Robot or LAWS?

A killer robot or lethal autonomous weapons system (LAWS) is a device that has sensors, cognition and actuators the combination of which is designed to kill enemy personnel and/or destroy enemy military objects. It is a robot or "autonomous weapon system" (AWS) designed to perform some potentially lethal role in fighting a war. The sensors of the LAWS might be radar, sonar, video cameras or other devices. Cognition is typically software but sometimes it is mechanical.

When analysing the functionality of robots it is typical to break it down into sensors, cognition and actuators. The sensors "perceive" or sense the world, cognition

S. Welsh (✉)
Department of Philosophy, University of Canterbury, Christchurch, New Zealand
e-mail: sean.welsh@pg.canterbury.ac.nz

M.I. Aldinhas Ferreira et al. (eds.), *A World with Robots*,
Intelligent Systems, Control and Automation: Science and Engineering 84,
DOI 10.1007/978-3-319-46667-5_13

"thinks" about what has been sensed and decides what to do, actuators "do" what has been decided. In the context of LAWS, the actuators will be lethal weapons such as chain guns, Hellfire missiles and laser-guided bombs such as described in Arkin (2009).

Some think the term "killer robot" objectionable. Heyns (2013) calls it "emotive". Lokhorst and van den Hoven (2011) say it is an "insidious rhetorical trick." Lokhorst and van den Hoven point out that robots in war could be engineered to do other things such as capture not kill and perform tasks involving humanitarian relief that would need them to be armed but not necessarily lethal. They could be wounding robots or robots that defend themselves against humans seeking to disrupt their missions with tasers or other non-lethal technologies. Thus they propose the term armed military robot instead of killer robot. Galliott (2015) speaks of military robots. The UN Special Rapporteur on extrajudicial, summary or arbitrary executions (Heyns 2013) has used the term lethal autonomous robot (LAR). The term used at the United Nations in the two Expert Meetings held by signatories to the CCW in May 2014 and April 2015 was lethal autonomous weapons system (LAWS).

The term "lethal autonomous weapons system" can cover weapons systems that have distributed architectures and that are actually fielded and under development. It is the term that has been adopted by the United Nations thus it is preferred for serious analysis. However if one wanted to discuss robots or weapons systems that have non-lethal military uses one might want to drop the L in LAWS and speak of AWS or military robots instead. In this paper my concern is with defining terms to regulate or ban lethal robots in a UN context so I stick to LAWS.

13.2 Proponents and Opponents

There is a range on views on LAWS. While LAWS should not be confused with remotely piloted vehicles (RPVs), if the network link used for telepiloting goes down a RPV such as the Predator B-1 will go "off the loop." Questions about RPVs naturally overlap with questions about LAWS. If the network link goes down (or is shot out) is the RPV that has effectively become autonomous permitted to failover to LAWS "off the loop" mode to defend itself? If not, why not?

One of the more barbed exchanges of diplomatic views at the two Expert Meetings were the statements of the United States delegation and the Pakistani delegation. The Americans, two years running, said RPVs should not be discussed at this forum because they are RPVs not LAWS (United States 2014, 2015). The Pakistanis, two years running, said, in the same forum, ban RPVs and ban LAWS (Pakistan 2014, 2015).

It is not my purpose in this paper to argue which view is correct or where the truth lies between these two positions. My aim is merely to define concepts you would need to draft "realistic regulation" to either ban or regulate LAWS (Galliott 2015).

Debates on lethal robots can be impassioned and pointed. One could say that the debate on "killer robots" is the "sharp end" of robot ethics. Lucas (2013) makes trenchant criticisms of both sides worth quoting at length:

> Much of this dispute between proponents and critics of enhanced machine autonomy is mired in a nearly hopeless kind of conceptual confusion and linguistic equivocation. Proponents of increasing machine autonomy, for their part, sometimes complicate the issues unnecessarily by invoking what turn out to be spurious concepts, like machine 'morality,' or by describing their proposals for an 'ethical governor' for lethally armed autonomous robots. They misleadingly describe autonomous combat weapon systems that would be empowered to make 'moral decisions and judgements,' and that would also (in principle at least) experience the machine equivalent of 'guilt' from sorties gone wrong, and 'learn' from those experiences. Consequentially, these proponents argue, lethally armed, autonomous military robots will be 'more ethical' and even 'more humane' than their human counterparts.
>
> Critics for their part worry needlessly about 'killer robots' run amok, as well as the presumptive moral inappropriateness of machines 'making decisions to kill humans,' or the lack of meaningful accountability for resulting 'war crimes' that might consequently be committed. The critics appear to envision cyborgs (like 'the Terminator') or the infamous intelligent computer 'HAL' (from Arthur C. Clarke's science fiction novel 2001: A Space Odyssey) in command on the bridge of a nuclear submarine, or 'R2D2' and 'C3P0,' fully weaponized and roaming the mountains of southern Afghanistan but unable to distinguish (without human supervision) between an enemy insurgent and a local shepherd.
>
> Both extremes are, frankly, preposterous.

Of course, when Lucas refers to an "ethical governor" he is referring primarily to Arkin (2009). Others arguing for LAWS include Anderson and Waxman (2013), Schmitt and Thurnher (2012) and Brooks (2015). As for the critics worried about "killer robots" he is referring to the Campaign to Stop Killer Robots and the various individuals and human rights organizations that support it. Papers arguing against LAWS include Sharkey (2009, 2010, 2012), Sparrow (2007, 2012), Asaro (2009), Matthias (2004, 2011), Article 36 (2013) and Heyns (2013).

The overall thrust of these papers is that robots should not decide to kill humans. However, decisions made by LAWS are complex. There are three critical phases of decision making in lethal action that can involve humans or robots or both.

The aim of this paper is to cut through the "conceptual confusion" and "linguistic equivocation" and come up with clear language and definitions fit for inclusion in a regulatory instrument such as a Protocol VI of the Convention of Certain Conventional Weapons (CCW). The notion of "meaningful human control" will be examined with a view to establishing whether such language should be part of an international binding treaty instrument.

13.3 LAWS Definition

LAWS are often described as in the loop, on the loop or off the loop (Scharre 2015). To these common terms I add two more types of LAWS and refer to them all by numbers. Type 0 denotes RPVs which have no robot on the lethal loop at all. Type

4 is for what I call "no human on the loop" type weapons where the robots have evolved "beyond the human loop" in terms of their adopted norms. It is a category in which one might put the likes of Skynet, the superintelligent malevolent computer in the Terminator movies. However, it could also include normative systems that use technologies such as machine learning (Alpaydin 2011), genetic algorithms (Liu 2009) and reinforcement learning (Sutton et al. 1998) to "create their own moral reality" (Boella and et al. 2008). Such technologies might be banned from the "ethical governor" of a LAWS i.e. the normative cognition of the system. However, they might be permitted in modules controlling tactics.

To illustrate, in a robot that plays the board game Monopoly, it would not be acceptable for the robot to decide to change the rules of the game. It would be acceptable for the robot to use machine learning and the like to decide whether or not to buy expensive, middle or cheap properties or not and when in the game to do so based on decisions of other players. These decisions one could call tactical. The decision about what to do when landing on "Go to Jail" is normative. When the robot player lands there it must move directly to Jail. This type of norm-determined decision cannot be subject to any rewriting by machine learning.

For military robots, the equivalents of the "rules of the game" in Monopoly are International Humanitarian Law (IHL) and such Rules of Engagement (RoE) as the operating power has defined through the chain of command. Tactics can be machine learned by "adaptive" robots perhaps but norms with respect to the legality of targeting decisions cannot.

Beside two extra classes of LAWS to the usual three of in, on and off the loop, I also add the key consideration as to what kind of agent defines targeting criteria. By define targeting criteria I mean the various functions involved in object recognition and proportionality calculations described in Arkin. Many discussions refer to the "critical functions" of "selecting and engaging targets" (ICRC 2014) but in the context of LAWS defining targets is just as "critical" a function as selecting and engaging targets. Selecting includes the functions of searching for, detecting, evaluating and tracking targets (Adams 2001 [2001]). Engaging means the lethal act of firing a missile or other munition at the target.

A third addition to common definitions is to clearly separate lethal autonomy from other kinds of robot autonomy. Lethal autonomy refers only to the three critical functions of defining targets, selecting targets and engaging targets. In simple terms this means defining what to attack, selecting what to attack on the basis of the definitions and actually attacking with the force of lethal actuators (e.g. Hellfire missiles or whatever the LAWS fires). Existing RPVs have many forms of non-lethal autonomy. For example, some can take off and land without a human hand on the joystick. This kind of autonomy is not lethal autonomy and does not form part of the definitions of LAWS offered here.

A fourth addition is the notion of final say. It may be that a human uses a computer or robotic agent to decide that a target should be selected. Thus there may be grey areas in the definitions where a human uses robot data and recommendations from robot agents to make decisions about selecting and engaging targets. In such "mixed agent" cases, the agent with final say in the decision determines the LAWS type.

Table 13.1 Types of LAWS

Type	Label	Critical functions
0	No robot	Human defines, selects and engages
1	Human in the loop	Human defines, robot selects, human confirms, robot engages
2	Human on the loop	Human defines, robot selects, human can abort, robot engages if no abort
3	Human off the loop	Human defines, robot selects, robot engages, human cannot abort
4	No human	Robot defines, selects and engages

For example, a weapons system's sensors and cognition might tag a tank as T-80 and note that this tank is an enemy tank. If the human looks at the screen and concurs with the robot select decision and advises another human to engage the tank (or presses the fire button himself), we can deem this human selects the target even though the human had robotic assistance because the human had "final say" before passing the decision on down the lethal loop. Conversely, if a human soldier in a trench during a sandstorm hears the sounds of tanks and asks a robot to check out the sound, if the robot using infra-red sensing, radar or acoustic signatures determines that the tank is an enemy tank and thus should be selected as a target, then the robot agent has final say in the select decision.

In summary, using these concepts, LAWS can be divided into five types (Table 13.1).

13.4 Type 0 LAWS

A Type 0 LAWS is an RPV such as the Predator B-1 drone. It is remotely piloted by human beings. Targeting criteria are defined by human beings. Humans select and humans engage the target. Robot cognition does not have final say in the critical functions of making decisions regarding defining, selecting and engaging the target. While the missile might be of the "fire and forget" variety and thus have autonomy in actually hitting the target once fired, the robot has no say in the decision to engage (i.e. to fire) (Fig. 13.1).

While some nations (e.g. Pakistan, Palestine) object to RPVs, their banning is not generally regarded as being within the purview of a proposed Protocol VI of the CCW to regulate or ban LAWS.

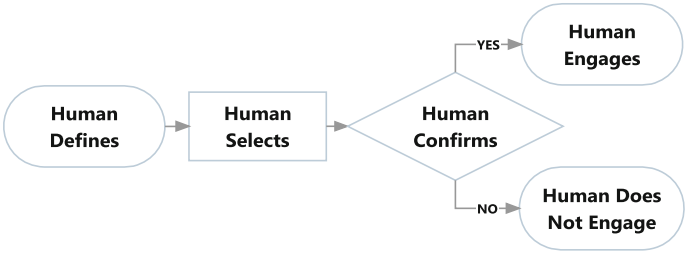


Fig. 13.1 No robot on the loop

13.5 Type 1 LAWS

A Type 1 LAWS is not really fully "autonomous" regarding its lethality in that a human still has to agree with the robot decision to select the target. In this kind of LAWS robots do not make up their own targeting criteria. The robot uses human defined criteria to select targets and calculate proportionality. The robot will then present via some suitable user interface, sufficient information for the human operators to agree or disagree with the robot decision. In this architecture, the human must press a button or make some positive act to confirm the kill decision. The robot then engages (Fig. 13.2).

The anti-missile system Patriot made by Raytheon is an example of a LAWS that can run in Type 1 mode. Patriot selects targets based on human defined criteria. It requires a human operator to confirm its decision. Patriot then engages the target with an anti-missile. During the 2003 Gulf War Patriot incorrectly engaged two friendly aircraft. However, overall, its fire/no-fire decisions were 99.99995 % correct (Defense Science Board 2005). (On the other hand if you restrict the count to fire decisions, 2/11 engagements or 18 % were fratricides.) The Samsung SGR-A1 "sentry robot" running in "normal mode" is another example of a Type 1 LAWS. The SGR-A1 selects a target and recommends a firing decision but waits for human confirmation before engaging.

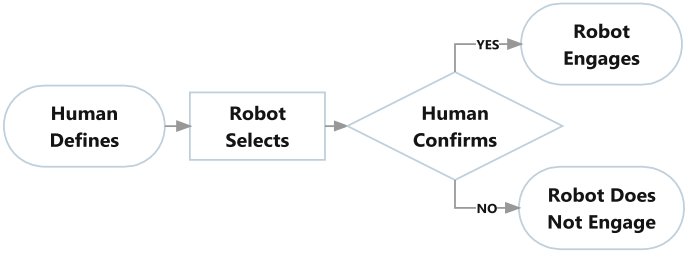


Fig. 13.2 Human "in the loop"

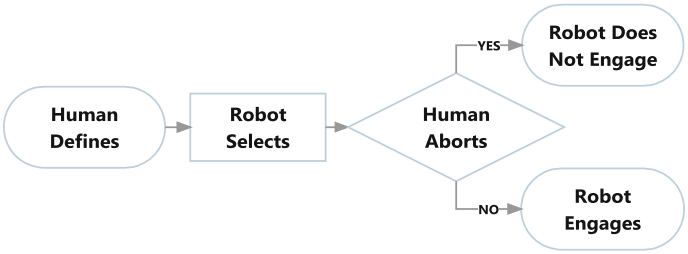


Fig. 13.3 Human "on the loop"

13.6 Type 2 LAWS

A Type 2 LAWS can kill without a human "finger on the trigger." The Type 1 and Type 0 LAWS require humans to make positive acts between the events of target selection and target engagement. They must press a button or pull a trigger. The Type 2 LAWS can kill without a human positive act. The human operator has the opportunity to abort the kill decision or override the robot's target selection. The presumption is that there is sufficient time and sufficient information for the decision of the human in this circumstance to be "meaningful." If the user interface is too cluttered or the human has insufficient time or training to abort in a timely fashion, then this "human control" is considered "meaningless" (Article 36, Killer Robots 2013). Similarly in the Type 1 architecture if the operator just hits "Confirm" without any consideration or thought in an immediate reaction like "Whack-a-Mole" in an arcade this too is "meaningless."

In the Type 2 architecture, if the human does nothing in a set time after target selection, then target engagement will occur without a human positive act. Thus, in the Type 2 decision tree, the robot can kill without any human approval other than the initial target selection criteria and the LAWS being turned on and sent into battle (Fig. 13.3).

Patriot can also run in Type 2 mode as can the Samsung SGR-A1 when set to "invasion mode" The drone in Arkin (2009) is explicitly designed to run in "on the loop" mode. The assumption in Arkin is that humans may want to override (abort) a robot fire decision or, indeed, that two humans may want to override (i.e. "force confirm") a robot no-fire decision.

13.7 Type 3 LAWS

Historically speaking, the oldest LAWS (if you include purely mechanical "autonomous" weapons in the definition) are the oldest forms of LAWS. The land "torpedoes" of the Confederacy that "blew to atoms" about a dozen of General Sherman's men as they stormed Fort McAllister near Savannah, Georgia in December

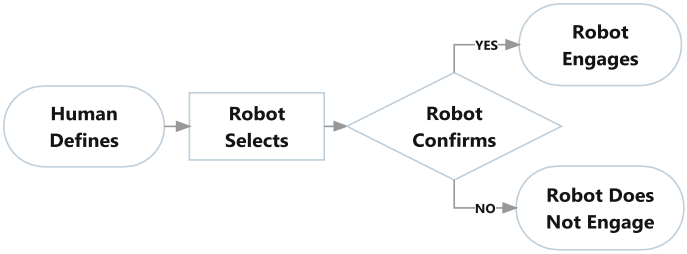


Fig. 13.4 Human "off the loop"

1864 (United States War Department 1891) fit into the definition of Type 3 LAWS given here. They were lethal weapons. They had "no human finger on the trigger" and so were "autonomous" (on the simple definition used in Arkin 2009).

The Type 3 LAWS can also kill without a human "finger on the trigger" and indeed if the robot engages in systematic mistargeting due to flawed target selection criteria or a buggy implementation of discrimination functionality it would not be possible for a human to hit an abort button to stop the robot. In the "off the loop" mode the robot takes its human defined targeting criteria and engages in its mission. The difference between the Type 2 and the Type 3 LAWS is very fine. If the human operator of the Type 2 LAWS is distracted, has insufficient time to process information or is killed the Type 2 LAWS becomes a de facto Type 3 LAWS (Fig. 13.4).

The likely military candidates for Type 3 LAWS are UUVs (submarine drones) which present intractable telepiloting challenges due to the bandwidth limitations of VLF and ELF radio transmission through sea water. Telepiloting submarines without tethers (underwater cables) is not presently viable. However, it is also the case that many existing weapons systems must function at decision speeds too fast for human review of select decisions. Often there is no time for humans to confirm or abort and thus the Type 1 and Type 2 LAWS architectures are not viable. As Adams noted back in 2001 in his visionary paper (Adams 2001 [2011]), the kinds of "lightning duels" involved in "close-in" weapons systems such as Phalanx occur at speeds too fast for the human brain to be a useful part of the decision tree. Adams robustly asserted that "warfare has begun to leave the human space" and that "we are faced with the prospect of equipment that not only does not require soldiers to operate it, but may be defeated if humans do attempt to exert control in any direct way."

The Type 0 to Type 3 LAWS all have human defined target selection criteria. The fundamental critical function of defining the enemy and what characterizes him remains in human hands. Robots "in the lethal loop" act as delegated agents following instructions in the form of making select and engage decisions based on targeting criteria that are human defined. At the 2015 Expert Meeting the Israeli delegation observed that the phrase "meaningful human control" was used in different ways by different delegations (Israel 2015). In the terms defined above, some thought it meant a ban on Type 2 or Type 3 LAWS. Others thought that it only meant a ban on Type 4. There are some of course (e.g. Pakistan) who would like to ban all types of LAWS, who think that even RPVs (Type 0 LAWS such as drones) are a moral aberration.

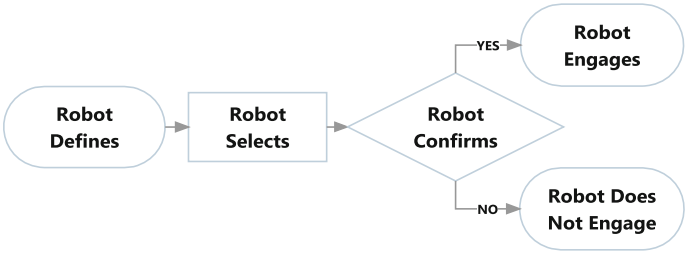


Fig. 13.5 No human on the loop

13.8 Type 4 LAWS

The Type 4 LAWS is a category where there is no longer any human input into the lethal loop. One could classify Skynet, the superintelligent robot of the Terminator movies, as a Type 4 LAWS. Some are seriously worried about the possibility of a "superintelligence" emerging that will take over the world. Such a prospect is a "global catastrophic risk" (Yudkowsky 2008). It seems reasonable to ban such robotic agents that are "beyond human control" (which may not be "superintelligent" but equipped with "machine learning") and that have "no humans on the loop" from being in control of lethal actuators. One would think that almost everybody would be in favour of a ban on Skynet having control of nuclear missiles (Fig. 13.5).

In terms of the positions stated at the 2015 Expert Meeting on LAWS. Even the British have indicated that they might be amenable to a ban on Type 4 LAWS. "Meaningful human control" on the British reading would be achieved by human definition of targeting criteria and human decisions to turn the LAWS on and send them into battle. Others maintain that "meaningful human control" requires specific human consideration of the circumstances of each lethal decision by a human. Such a position would imply only Type 0 and Type 1 LAWS are normatively acceptable.

It is not my purpose here to adjudicate where the regulatory lines should be drawn. Merely to detail what the options are and to provide clear definitions that could be used to draw such lines.

13.9 Meaningful Human Control

Adams (2001 [2011]) discusses the notion of meaningful human control. "If the problem is how to maintain meaningful human control of autonomous warfighting systems, no good solution presents itself. One answer, of course, is to simply accept a slower information-processing rate as the price of keeping humans in the military decision business. The problem is that some adversary will inevitably decide that the way to defeat the human-centric systems is to attack it with systems that are not so limited."

This is the kernel of the problem. In IHL terms it comes down to the fundamental tension between the principle of military necessity (winning the war) versus the principle of humanity (the "requirements of the public conscience"). Winning wars involves acts of taking lives and destruction of valued property. Thus there are tensions between those pushing for the acceptance of a slower information-processing rate (i.e. Type 1 LAWS) and those pushing for the most effective military solution (i.e. a Type 3 LAWS) who bluntly assert in a fight between Type 1 and Type 3, Type 1 will lose.

At the 2015 Expert Meeting, the UK opposed a pre-emptive ban on LAWS (United Kingdom 2015). As the British have a permanent seat on the UN Security Council, this is a major blow for the Campaign to Stop Killer Robots. The British are in alliances with many of the world's major powers especially those that lead in robotics. These allies might come to support the British position.

It is clear that the British are committed to major investments in autonomous warfighting capabilities, notably in the form of Taranis, the combat drone developed by British firm BAE Systems. Interestingly, BAE Systems are on the consortium that is developing the F-35 Lightning, commonly said to be the last human-piloted fighter the US will ever build. Sooner or later there will be a dogfight between Taranis and Lightning. It will be interesting to see whether the Garry Kasparov type solution or the Deep Blue type solution will prevail in air war.

While Taranis is being publically advertised as a Type 1 LAWS, given the prevailing military opinion that increased autonomy will become a military necessity (United States Air Force 2009) because aircraft that operate at split second speed not achievable with satellite-based network lag will have critical tactical advantage over those that do not and further such aircraft will be less vulnerable to communications disruptions and cyberwarfare, it would seem obvious that Taranis will evolve into a Type 3 LAWS. Hence the British move to block a "comprehensive and pre-emptive ban" on LAWS via a Protocol VI of the CCW.

13.10 Meaningful Human Control of Individual Attacks

Article 36, by contrast, seeks to defend a line of "meaningful human control of individual attacks" (Article 36, 2013). This implies a Type 1 LAWS decision tree.

There are definitional questions as to what comprises an "attack." If a Samsung SGR-A1 detects a North Korean infantry battalion of 600 men charging across the DMZ does the human operator have to approve 600 kill decisions and hit the Confirm button 600 times or will it be acceptable to switch to "invasion mode" and not hit the Abort button while the robot fires?

It is unlikely that the military will accept that a human has to hit the Confirm button 600 times in this situation. Thus holding the line at Type 1 LAWS will be difficult unless it is admitted that sets of targeting decisions can be made.

Besides the question of a single normative decision that applies to a set of humans is acceptable in terms of meaningful human control of individual attacks, there is

the related question as to whether a class of moral decisions can be delegated from humans to robots. This is the heart of the matter.

No one these days worries about "meaningful human control" of speeding cameras and demands that there must be a "human in the loop" for such machines. No one doubts they are more accurate and less biased than humans. In the case of a speeding camera a class of normative decisions that harm humans (fining them for speeding) has been delegated to robots by humans.

That said, traffic law is one thing and IHL another. Speeding tickets can be reviewed by affected humans in court. If, by contrast, you are hit by a Hellfire missile instead of a speeding ticket, you are unable to seek redress.

There are vigorous debates as to whether robots can meet the requirements of IHL. The questions as to whether robots can comply with principles such as discrimination and proportionality are subject to empirical confirmation and refutation. Such questions will be resolved by existing Article 36 processes of weapons testing and review. Given the continuing advances in AI, it is possible that such principles will eventually be programmable. According to Russell (2015) face and gait recognition by AI is now at "superhuman" levels. While he was cautious to make clear this did not mean that robots could discriminate combatants in cluttered environments, it does mean that if video of a suspect terrorist is available then robots will be able to recognize him in a crowd better than humans. Technical claims that robots cannot discriminate and that robots cannot make proportionality calculations are vulnerable to empirical refutation as technology advances.

Even so there remain more fundamental arguments against using robots to kill humans. Even if all the technical premises are granted, that robots can discriminate, calculate proportionality and that mechanisms of accountability can be devised, it is still possible to make a purely normative claim that robots should not be used to kill humans because tasking robots with human destruction is an affront to a fundamental human right to dignity (Sparrow 2012; Heyns 2015; Lin 2015).

13.11 Conclusion

Adams (2001 [2001]) predicted that "future generations may come to regard tactical warfare as properly the business of machines and not appropriate for people at all. Humans may retain control at the highest levels, making strategic decisions about where and when to strike and, most important, the overall objectives of a conflict. But even these will increasingly be informed by automated information systems. Direct human participation in warfare is likely to be rare. Instead, the human role will take other form—strategic direction perhaps, or at the very extreme, perhaps no more than the policy decision whether to enter hostilities or not."

Adams thought military robots would lead to a decline in human decision-making, though one could argue AI and Robotics will augment human cognition not lead to its decline. It is certainly the case that human cognition is being displaced (Carr 2015).

The evolution of military robotics will provide deep challenges for those working in robot ethics.

In this paper, it has not been my purpose to recommend the banning or regulation of any Type of LAWS, merely to define terms that might be used in a treaty. The phrase "meaningful human control" can have a wide range of meanings and thus is too vague for inclusion in regulation. The status quo is that no LAWS are illegal provided they can be operated in conformance with IHL. IHL does regulate existing LAWS and no expert disputes that it applies to present and future LAWS (ICRC 2014). Thus it is not the case that there is a regulatory vacuum.

Acknowledgements Jack Copeland, Michael-John Turp, Walter Guttmann, Christoph Bartneck, Ron Arkin, Paul Scharre, Michael Horowitz.

References

Adams T (2001 [2011]) Future warfare and the decline of human decisionmaking. Parameters 41(4):1–15

Alpaydin E (2011) Machine learning. Wiley Interdiscip Rev: Comput Stat 3(3):195–203

Anderson K, Waxman M (2013) Law and ethics for autonomous weapon systems: why a ban won't work and how the laws of war can. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2250126. Accessed 12 Feb 2015

Arkin R (2009) Governing lethal behaviour in autonomous robots. CRC Press, Boca Rouge

Article 36 (2013) Killer Robots: UK government policy on fully autonomous weapons. http://www.article36.org/wp-content/uploads/2013/04/Policy_Paper1.pdf

Asaro P (2009) Modelling the moral user. IEEE Technol Soc Mag 28(1):20–24

Boella G et al (2008) Introduction to the special issue on normative multi-agent systems. Auton Agents Multi-Agent Syst 17(1):1–10

Brooks R (2015) In defense of killer robots. Foreign Policy. http://foreignpolicy.com/2015/05/18/in-defense-of-killer-robots/. Accessed 27 May 2015

Carr N (2015) The glass cage. Random House, London

Defense Science Board (2005) Patriot system performance. http://www.acq.osd.mil/dsb/reports/ADA435837.pdf. Accessed 18 Feb 2015

Galliott J (2015) Military robots: mapping the moral landscape. Ashgate, Farnham, UK

Heyns C (2013) Report of the special Rapporteur on extrajudicial, summary or arbitrary executions. http://www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Session23/A-HRC-23-47_en.pdf. Accessed 16 Feb 2015

Heyns C (2015) Comments by Christof Heyns, United Nations Special Rapporteur on extrajudicial, summary or arbitrary executions. http://unog.ch/80256EDD006B8954/(httpAssets)/1869331AFF45728BC1257E2D0050EFE0/38file/2015_LAWS_MX_Heyns_Transcript.pdf. Accessed 10 Oct 2016

ICRC (2014) Statement of the International Committee of the Red Cross. http://unog.ch/80256EDD006B8954/(httpAssets)/C99C06D328117A11C1257CD7005D8753/38file/ICRC_MX_LAWS_2014.pdf. Accessed 10 Oct 2016

Israel (2015) Intervention by Israeli Delegation in Characteristics of LAWS (Part II). http://unog.ch/80256EDD006B8954/(httpAssets)/AB30BF0E02AA39EAC1257E29004769F3/38file/2015_LAWS_MX_Israel_characteristics.pdf. Accessed 10 Oct 2016

Lin P (2015) The right to life and the Martens Clause. http://unog.ch/80256EDD006B8954/(httpAssets)/2B52D16262272AE2C1257E2900419C50/38file/24+Patrick+Lin_Patrick+SS.pdf. Accessed 10 October 2016

Liu B (2009) Genetic algorithms. In: Theory and Practice of uncertain programming. Springer, Berlin, pp 9–17
Lokhorst G, van den Hoven J (2011) Responsibility for military robots. In: Lin P, Abney K, Bekey G (eds) Robot ethics: the ethical and social implications of robotics. MIT Press, Cambridge, MA
Lucas G (2013) Engineering, ethics and industry: the moral challenges of lethal autonomy. In: Strawser B (ed) Killing by remote control: the ethics of an unmanned military. OUP, New York, pp 211–228
Matthias A (2004) The responsibility gap: ascribing responsibility for the actions of learning automata. Ethics Inf Technol 6(3):165–183
Matthias A (2011) Is the concept of an ethical governor philosophically sound? STAND 338
Pakistan (2014) Statement by Ambassador Zamir Akram. http://unog.ch/80256EDD006B8954/(httpAssets)/D0326935FBB0FB7EC1257CE6005465A6/38file/Pakistan+LAWS+2014.pdf. Accessed 10 Oct 2016
Pakistan (2015) Statement by Irfan Mahmood Bokhari. http://unog.ch/80256EDD006B8954/(httpAssets)/C6F268A1B1D7B80BC1257E26005E33E4/38file/Statement+on+LAWS+-+CCW+Informal+Meeting+of+Experts+April+2015.pdf. Accessed 10 Oct 2016
Russell S (2015) Artificial intelligence: implications for autonomous weapons. http://unog.ch/80256EDD006B8954/(httpAssets)/36AF841749DE9819C1257E2F0033554B/38file/2015_LAWS_MX_Russell+bis.pdf. Accessed 10 Oct 2016
Scharre P (2015) Presentation at the united nations convention on certain conventional weapons. http://unog.ch/80256EDD006B8954/(httpAssets)/98B8F054634E0C7EC1257E2F005759B0/38file/Scharre+presentation+text.pdf. Accessed 10 Oct 2016
Schmitt M, Thurnher J (2012) Out of the loop: autonomous weapon systems and the law of armed conflict. Harv Natl Secur J 4:231
Sharkey N (2009) Death strikes from the sky: the calculus of proportionality. IEEE Technol Soc Mag 28(1):16–19
Sharkey N (2010) Saying 'no!' to lethal autonomous targeting. J Mil Ethics 9(4):369–383
Sharkey N (2012) Killing made easy: from joysticks to politics. In: Lin P, Abney K, Bekey G (eds) Robot ethics: the social and ethical implications of robotics. MIT Press, Cambridge, MA, pp 111–128
Sparrow R (2007) Killer robots. J Appl Philos 24(1):62–77
Sparrow R (2012) Can machines be people? Reflections on the turing triage test. In: Lin P, Abney K, Bekey G (eds) Robot ethics: the ethical and social implications of robotics. MIT Press, Cambridge, MA, pp 301–316
Sutton R, Andrew G, Barto A (1998) Introduction to reinforcement learning. MIT Press
United Kingdom (2015) Statement to the informal meeting of experts on lethal autonomous weapons systems. http://unog.ch/80256EDD006B8954/(httpAssets)/1CBF996AF7AD10E2C1257E260060318A/38file/2015_LAWS_MX_United+Kingdom.pdf. Accessed 10 Oct 2016
United States (2014) Opening statement. http://unog.ch/80256EDD006B8954/(httpAssets)/E7CB7B95715BFEB4C1257CD7005DCD54/38file/USA_MX_LAWS_2014.pdf. Accessed 10 Oct 2016
United States (2015) Opening statement. http://unog.ch/80256EDD006B8954/(httpAssets)/8B33A1CDBE80EC60C1257E2800275E56/38file/2015_LAWS_MX_USA+bis.pdf. Accessed 10 Oct 2016
United States Air Force (2009) Unmanned aircraft systems flight plan 2009-2047. http://fas.org/irp/program/collect/uas_2009.pdf. Accessed 13 May 2015
United States War Department (1891) Official Report No. 21. The war of the rebellion: a compilation of the official records of the Union and Confederate armies, vol 16. Government Printing Office, Washington
Yudkowsky E (2008) Artificial intelligence as a positive and negative factor in global risk. In: Bostrom N, Ćirković M (eds) Global catastrophic risks. Oxford University Press, New York, pp 308–345

Chapter 14
Safety Issues of the Portuguese Military Remotely Piloted Aircraft Systems

Delfim Dores, Ana Baltazar, Teresa Cabral, Isabel Machado and Paula Gonçalves

Abstract This paper provides an overview of the safety issues of the Portuguese Military Remotely Piloted Aircraft Systems (RPAS), namely the human error, integration into regulated common national airspace (considering the rules of air) and the airworthiness certification aspects. The *Autoridade Aeronáutica Nacional* (AAN) safety requirements for the RPAS airworthiness certification and operation authorization processes are presented. This paper also brings out the safety assessment methodology by addressing its application to the Antex-X02 RPAS, a platform under development by the Portuguese Air Force Academy. The result of the safety assessment has contributed to obtain a Permit to Fly, a clearance issued by the AAN.

Keywords Airworthiness · Military · Risk · RPAS · Safety

14.1 Introduction

In general, aviation safety depends on the error minimization in all the phases of the aviation system: design, production, maintenance, flight training and operations. To achieve the required levels of aviation safety, different data from different sources needs to be constantly considered and the existence of an aviation policy must be seen by the states as an individual obligation.

This paper is focused only on RPAS, mainly on the design safety aspects. In fact, this is a paramount issue to military, as well as the civil Authorities, because of the need to integrate the RPAS in the national Airspace System (DeGarmo 2004; Evans

D. Dores · A. Baltazar (✉) · I. Machado · P. Gonçalves
Instituto Universitário Militar, CISDI Researcher, Lisbon, Portugal
e-mail: arbaltazar73@gmail.com

T. Cabral
Academia da Força Aérea Portuguesa, CIAFA Researcher, Lisbon, Portugal

M.I. Aldinhas Ferreira et al. (eds.), *A World with Robots*,
Intelligent Systems, Control and Automation: Science and Engineering 84,
DOI 10.1007/978-3-319-46667-5_14

and Nicholson 2007). Civil and military RPAS are different, however, both share the same airspace and some safety requirements (e.g. not to increase the risk to third parties).

Current study is aligned with EASA concerns as presented in the recently published "Concept of Operations for Drones—A risk based approach to regulation of unmanned aircraft" (European Aviation Safety Agency 2015). This concept was developed taking into consideration two main goals: the integration of RPAS in the aviation system in a safe and proportionate manner; and the growth of European industry in order to create new employment. These goals will be achieved by developing research on: Detect and Avoid, Airspace and Airport access, Command and Control Communications, Human Factors, Contingency, Security, and Autonomy of the RPAS (European Aviation Safety Agency 2015).

This article is a summary of a project being developed by two research centres (CISDI and CIAFA), entitled "The error in military operations: the Portuguese Air Force case". This research is focused on military aircraft (e.g. RPAS), in particular on the study of different errors approaches: organizational (related to management) and technical (related to engineering). The aim of the project is to find out how errors can be avoided and what lessons can be obtained from them in order to increase the safety level. The Antex-X02 is herein presented as a case study.

The study follows a top down approach starting with generic safety issues and then focusing on the aircraft operation, including the integration into non-segregated airspace, airworthiness certification, and verification/demonstration of compliance with the established standards to assure an acceptable level of safety (Evans and Nicholson 2007).

14.2 Safety

The International Civil Aviation Organization (ICAO) defines Safety as "the state in which the risk of harm to persons or property damage is reduced to, and maintained at or below, an acceptable level through a continuing process of hazard identification and risk management" (International Civil Aviation Organization 2013b). The accidents that have happened remind us that even though the technology has evolved, it has not been able to stop them from happening. To mitigate these accidents, we need to study the risk areas (e.g. training, software, hardware, rules, standards, laws).

Focusing on the development stage, we find different ways to validate and verify the process that will permit the aircraft to fly safely and reliably. According to DeGarmo (2004) reliability could be improved by developing the integrity of components and systems and/or by developing systems redundancies. Reliability being the probability of a component or system to perform a required function for a period of time, when used in stipulated operating conditions (Ebeling 2005).

Furthermore, all the maintenance tasks, performed during the life cycle of the aircraft by certifying staff, are a key role to ensure the reliability of all parts of the system (i.e. vehicle, ground station, communication equipment, and command and

control link). In terms of maintenance occurrences, Dhillon (2009) states that they can be originated by several causes like: complex maintenance tasks, poorly written maintenance procedures, fatigued maintenance personnel, inadequate training and experience, poor work layout, poor work environment, improper work tools, poor equipment design, and outdated maintenance models. This is the type of error that we could call "latent error" because the consequence is the result of an action or decision taken much earlier than the occurrence (Reason 1997), making it difficult to understand why it happened and, consequently, difficult for one to learn with the process.

In the aeronautical industry, people do tend to associate the "human factors" to cockpit and/or to crew (i.e. active error, where the error consequence is immediate and the person is present). In this case, we do not have only pilots (or crew) on board but many different people (e.g. engineers, technicians) involved in the process (e.g. design, function, performance), who must be studied in order to understand the origin of human error. When we separate the man from the machine, we change the way of doing things because we do not have access to some data (e.g. visual data), increasing the dependency on the hardware and software. The different interfaces involved in this type of systems justify different approaches to the safety issue. That is not the aim of this paper.

As proposed by the Civil Aviation Authority (CAA) (2002), the human factors "are concerned with optimizing the relationship between people and their activities by the systematic application of the human sciences integrated in the framework of engineering systems". CAA defends that one error assigned to a human can be originated by a deficient system design, by a lack of training, by wrong procedures or by mistakes and/or omissions in manuals.

When dealing with RPAS flight, another safety issue emerges: operating in a common airspace, with high intensity air traffic flows. Once again, part of the solution relies on technology (e.g. detection) and in the development of regulatory framework. The aim of these is to achieve an equivalent level of safety (comparing with manned aircraft), starting with the airworthiness certification requirements. This is also aligned with European Aviation Safety Agency when it states that: "Regulatory airworthiness standards should be set to be no less demanding than those currently applied to comparable manned aircraft nor should they penalise UAS by requiring compliance with higher standards simply because technology permit" (European Aviation Safety Agency 2005).

14.3 RPAS Integration in Airspace

According to ICAO, air traffic will double in the next 15 years (International Civil Aviation Organization 2013a). To prepare for this air traffic increase, States and involved geographical regions should focus on improving safety procedures in synchronisation with all stakeholders to embrace the air transport sector's expansion. It is fundamental to continually improve global aviation safety, to ensure that air transport

is developed in a sustainable economic, social, and technologic way. ICAO, States, and aviation stakeholders consider aviation safety and air navigation the means and targets to efficiently manage air traffic growth. Consequently, they have already published the Global Aviation Safety Plan, from 2014 to 2016, and the Air Navigation Global Plan, with the aim of at least maintaining the present safety outcomes (International Civil Aviation Organization 2013a). The RPAS are included in this growth.

In line with the existing ICAO airspace rules for manned aircraft, the integration of RPAS in the common airspace shall always have in mind that the rules of air are standards, established to ensure a minimum level of safety.

There are already some regulations applicable to the integration of RPAS in airspace. Nevertheless, RPAS are not yet completely regulated, whereby each national authority must perform a case by case analysis, framed by the international and European policies, such as ICAO, Eurocontrol, EASA, and FAA's practices. For military RPAS specifically, besides the due regard for civilian policies and rules, it is necessary to foreseen interoperability between European Defence Agency, North Atlantic Treaty Organization (NATO) and national operational directives.

In order to identify a RPAS level of safety and conclude about its integration in ICAO airspace, two main variables should be analysed: applicable rules of the air and airworthiness certification requirements.

According to ICAO Annex 2, the rules of the air divide the airspace in seven classes, identified as A, B, C, D, E, F, and G, organized according with the flight rules (IFR/VFR), type of air traffic services, rules of operations (velocity limitation, Air Traffic Control (ATC) clearance, etc.), and separation responsibility. In addition to these classes, there are also restricted areas, temporary reserved areas, dangerous areas, prohibited areas, civil or military segregated airspace, Control Zones (CTR) and Control Areas (CTA), as Terminal Control Area (TMA) and Airway (AWY). The five first classes (A, B, C, D and E) are controlled airspace and the last two classes belong to uncontrolled airspace (F and G). Figure 14.1 presents the differences between the ICAO airspace classes.

ICAO considers two types of traffic: the General Air Traffic (GAT), to which all in force rules and obligations are applicable (civil use consists mainly of commercial airline movements); and the Operational Air Traffic (OAT), for flights that do not comply with GAT rules and obligations but have specific rules and procedures defined by the competent national authority (most OAT are operated by military aviation).

The baseline principle to integrate a RPAS in an airspace class shall be the safety of persons (crew, passengers, and third parties), aircraft, facilities and goods, whenever the RPAS is in flight or on the ground. In addition, other issues should also be considered such as privacy, data protection and ethics within the operation approval (European Commission 2014).

The required level of safety must be demonstrated by both the Design and the Operating Organizations to the competent authority. This demonstration comprises RPAS functions and capabilities: specifications, on-board sensors, communication, navigation, command and control links, sense and avoid systems, ATM integration, autonomy, human factors, airspace and airports access, security, environmental, spectrum of operations and limitations, among others.

In particular, the ethics issues should be included in the military operational assessment taking into consideration the discussion presented by Cortright et al. (2015). Nevertheless, this assessment depends upon the established acceptable conditions for the specific operations.

The ATC clearance for a controlled airspace class is issued for RPAS complying with the applicable rules, which will depend on the RPAS' demonstrated level of safety. If such demonstration has been already accepted by a competent authority (issuing either an Airworthiness Certificate or a Permit to Fly), the ATC clearance can be obtained through a more straight process. Nevertheless, if the demonstration is not complete, an ATC clearance can also be obtained through a more analytical and detailed process. This process may result in the application of necessary restrictions to ensure the required level of safety (i.e. restricted to a specific airspace class and/or area). Taking into account that a complete regulatory framework for the RPAS is not yet available, nowadays, to allow RPAS operations, it is necessary to segregate an airspace area and to ensure a previous coordination amongst Operator, Air Navigation Services Providers (ANSP), and aerodrome(s), in order to maintain the safety of manned aviation.

The necessary coordination between RPAS Operator and the ATC must be put in place to guarantee the correct interface between both (essentially for CNS, but also for normal and emergency procedures to maintain manned aviation safe). The coordination with the ANSP is important to confirm the amount and format of the exchanged data, to ensure correct interface and compliance with sending and receiving requirements and its understandability for the following: Air Traffic Management services, aeronautical services, meteorological services, search and rescue, and communication, navigation and surveillance. The aerodrome coordination is important to ensure the safe operation for the RPAS and other aircraft on ground and on take-off phase,

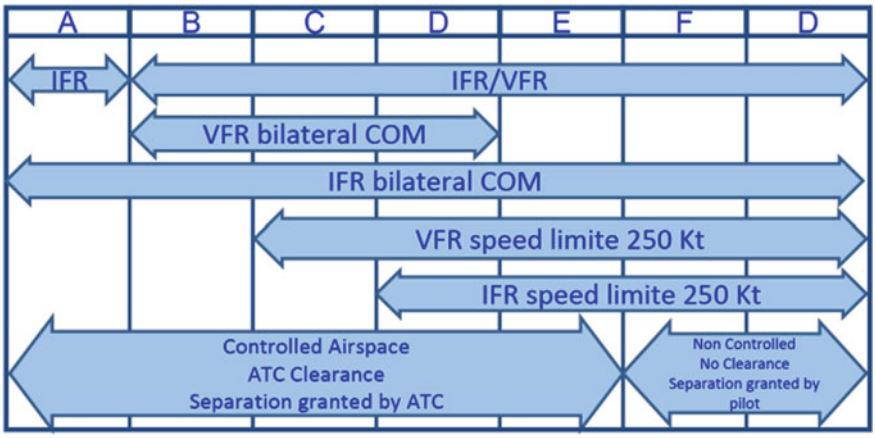


Fig. 14.1 ICAO airspace classification

depending on the existing type and density of air and ground traffic (International Civil Aviation Organization 2011).

Meanwhile, to perform a RPAS operation integrated with manned aviation in non-segregated airspace, it is necessary to identify the open issues and face the challenges of progress and development for all stakeholders. In this environment, the operation of RPAS in non-segregated airspace should be granted in a case by case analysis, within the ATC Clearance process or made by the competent authority accordingly with its published regulation Center for Advanced Aviation System Development (2012), and always depending upon the desired airspace class.

While technology is still being developed, the integration of RPAS in non-segregated airspace should be done in a progressive and proportionate way, jointly developed between the commercial needs, the industries interests, and the development of the regulatory framework. Accordingly with today's experience, this integration will probably occur on a step by step approach of shared development, and may encompass the following phases: RPAS in development can fly only in segregated airspace; RPAS can fly in non-segregated airspace, as OAT; and RPAS can fly in non-segregated airspace, as GAT. In these three phases, additional measures and assessments will possibly be established in order to ensure the required level of safety, while there is not sufficient confidence on the RPAS airworthiness.

At the moment, the Portuguese military RPAS are only integrated in segregated airspace. The Autoridade Aeronáutica Nacional (AAN[1]) requires a mitigation action plan, to ensure a minimal level of safety. In particular, depending on the intent of use and mission, the segregated airspace may occur over a military facility (e.g. Air Force Base), in a military segregated airspace area, or even in another common airspace area.

14.4 RPAS Safety

Safety is the key requirement for airworthiness; therefore, safety considerations must be present in the early stage of any RPAS design. It allows for the timely identification of hazards and their mitigation methodologies, avoiding the increase of RPAS costs due to rework and redesign tasks during the design and production processes (Evans and Nicholson 2007).

RPAS designers are required by airworthiness authorities to demonstrate that vehicles are compliant with an acceptable level of safety, when they apply for a special permit to fly or special airworthiness certificate (Evans and Nicholson 2007). To apply for a RPAS special permit to fly, AAN requires the development of an airworthiness certification for RPAS above 20 kg, or an operation authorization process for RPAS below 20 kg (Autoridade Aeronáutica Nacional 2013).

According to AAN Circular No. 1/13, each applicant shall present a RPAS safety case which consists of substantiation and a statement that the RPAS is safe to operate

[1] AAN is the Portuguese Aviation Authority which oversights the aviation matters in defense domain.

according to an appropriate body of evidences. The substation of the safety case shall include a safety evaluation based on the response of a safety checklist and on a safety assessment / risk analysis of the RPAS. The safety checklist enclosed in the FAA Order 8130.34 is considered as an acceptable means of compliance for the safety analysis (Autoridade Aeronáutica Nacional 2013).

For the safety assessment/risk analysis, it is considered as an acceptable means of compliance one or a combination of different tools such as Failure Mode Effects and Criticality Analysis (FMECA), Functional Hazard Analysis (FHA), Fault Tree Analysis (FTA), Dependence Diagram (DD), Failure Mode And Effect Analysis (FMEA) according to MIL-STD-1629, SAE ARP 4761, FAA AC 23.1309 or equivalent standards (Autoridade Aeronáutica Nacional 2013).

The safety assessment and risk analysis also includes evaluation of the area of operations, as well as the identification, assessment and control of operational hazards. Safety assessment is a widespread methodology for manned aircraft developers, as well as, an acceptable means of compliance for airworthiness authorities (Evans and Nicholson 2007). To perform a safety assessment for an RPAS, it is necessary to address aspects other than the aircraft itself. It should include control ground station, data links, mission planning, interoperability with ATC and other aircraft, operation environments, mission types, operator's competences and their procedures, level of autonomy and its predictability, emergencies and abnormal flight conditions (Evans and Nicholson 2007).

For RPAS, the safety assessment requirement is presented in various airworthiness military standards such as STANAG 4671, STANAG 4702 and STANAG 4703. Such standards address the airworthiness requirements for RPAS which are intended to operate in non-segregated airspace. STANAG 4671 defines the airworthiness requirements for fixed wing UAV (unmanned aerial vehicle) systems with a maximum take-off weight between 150 and 20,000 kg (NATO Standardization Agency 2009), whereas the STANAG 4703 defines the airworthiness requirements for fixed wing UAV systems with impact energy above 66 joules and a maximum take-off weight below 150 kg. STANAG 4702 describes the airworthiness requirements for rotor wing UAV systems with a maximum take-off weight between 150 and 3,175 kg (NATO Standardization Agency 2014b). The minimum acceptable level of safety for UAV equipment, systems and their installations is defined in the USAR.1309, USAR-RW.1309 requirements and associated AMC 1309 for STANAG 4671 and STANAG 4702, respectively, whereas for STANAG 4703, it is defined in the ER.1.3.3 requirement and its Annex G. Such requirements present a risk reference system which is a combination of severity and probability reference systems. The minimum acceptable level of safety for UAV equipment, systems and their installations are shown through the risk reference system. The risk reference system presented in STANAG 4671 and STANAG 4702 is illustrated in Table 14.1. The reference risk system specified in STANAG 4703 is a function of the cumulative probability for catastrophic event and depends on the UAV maximum take-off weight (MTOW). The risk reference system described in STANAG 4703 is presented in Table 14.2.

The safety methodologies used to perform a safety assessment are presented in MIL-STD-882 and in SAE ARP 4761 (NATO Standardization Agency 2009,

Table 14.1 Risk reference system, NATO Standardization Agency (2009, 2014a)

Probability (P)	Catastrophic	Hazardous	Major	Minor	No safety effect
Frequent $P > 10^{-3}/h$	U	U	U	U	A
Probable $10^{-3}/h \leq P > 10^{-4}/h$	U	U	U	A	A
Remote $10^{-4}/h \leq P > 10^{-5}/h$	U	U	A	A	A
Extremelv remote $10^{-5}/h \leq P > 10^{-6}/h$	U	A	A	A	A
Extremely improbable $10^{-6}/h \leq P$	A	A	A	A	A

A—Acceptable; U—Unacceptable

Table 14.2 Risk reference system, NATO Standardization Agency (2014b)

Probability (P)	Catastrophic	Hazardous	Major	Minor
Frequent (P_A) PA $> 1000 \times P_E$	U	U	U	U
Probable (P_B) $100 \times P_E < P_B \leq 1000 \times P_E$	U	U	U	A
Remote (P_C) $10 \times P_E < P_C \leq 100 \times P_E$	U	U	A	A
Extremely remote (P_D) $P_E < P_D \leq 10 \times P_E$	U	A	A	A
Extremely improbable (P_E) $P_E \leq \frac{P_{CUM\_CAT}}{n°CAT \text{ failure conditions}}$	A	A	A	A
MTOW $<$ 15Kg Cumulative probability		$P_{CUM\_CAT} = 10^{-4}/h$		
15 Kg $<$ MTOW $<$ 150Kg Cumulative probability		$P_{CUM\_CAT} = 0,0015/MTOW/h$		

A—Acceptable; U—Unacceptable

2014a, b). The MIL-STD-882 includes systematic hazard identification, risk analysis and risk management Department of Defense United States of America (2012), whereas, the SAE ARP 4761 consists of an iterative method through FHA, Preliminary System Safety Assessment (PSSA), System Safety Assessment (SSA) at the UAV system and subsystem levels (Society of Automotive Engineers 1996).

The SAE ARP 4761 safety system concept is based on the definition safety objectives at FHA and their validation and verification in PSSA and SSA, respectively. This concept is illustrated in Fig. 14.2.

The FHA starts with the conceptual design phase and is a qualitative process, which consists on the identification of all functions of the UAV systems and its interfaces, identification and description of the failure conditions associated with the identified functions, and determination of the effects and severity of such failure conditions (NATO Standardization Agency 2014b). The definition of the FHA safety objectives is based on the classification of the failure conditions. It is a top-down approach and its outputs are the inputs for the PSSA (Society of Automotive Engineers 1996). The PSSA is associated with the design definition stage. It can be qualitative or quantitative and its purpose is to examine the proposed system architecture, in order to validate the safety of such design, to identify safety requirements for the components of that architecture and to determine if the proposed system

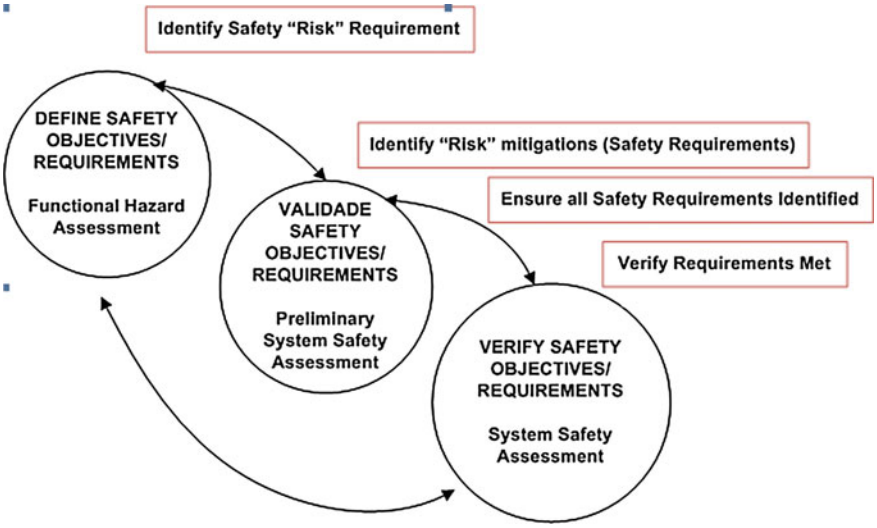


Fig. 14.2 SAE ARP 4761 safety system concept, Jones and Melding (2014)

architecture can meet the safety objectives defined in the FHA (Pumfrey et al. 1999; Society of Automotive Engineers 1996). The SSA is a bottom up method to verify that the safety objectives from FHA and derived safety requirements from PSSA are met in the implemented system (Society of Automotive Engineers 1996).

14.5 Aircraft Functional Hazard Assessment for the ANTEX-X02 Extended

This analysis comprises the FHA of the functions of the Antex-X02 Extended RPAS, presented in Fig. 14.3. The operation of all systems employed to accomplish the functions of the aircraft were considered. The Antex-X02 Extended RPAS is a Reconnaissance and Surveillance aircraft, equipped with an internal combustion engine, maximum speed of 89,5 mph, range up to 54 nm, and an average flight endurance of six hours.

In order to perform the FHA, the top-level RPAS functions are used as: aviate (fly the plane), navigate (fly it in the right direction), communicate (state your condition or intentions to other people) and mitigate hazards. Each of these functions comprises sub functions, as presented in Fig. 14.4 (National Aeronautics and Space Administration 2007; Society of Automotive Engineers 1996).

Aviate consists of activities such as flight control, ground control, command and control inputs and sub-systems control. Navigate comprises activities related to the management and following of a trajectory. In practice, communicate warrants the

Fig. 14.3 Antex-X02 extended RPAS

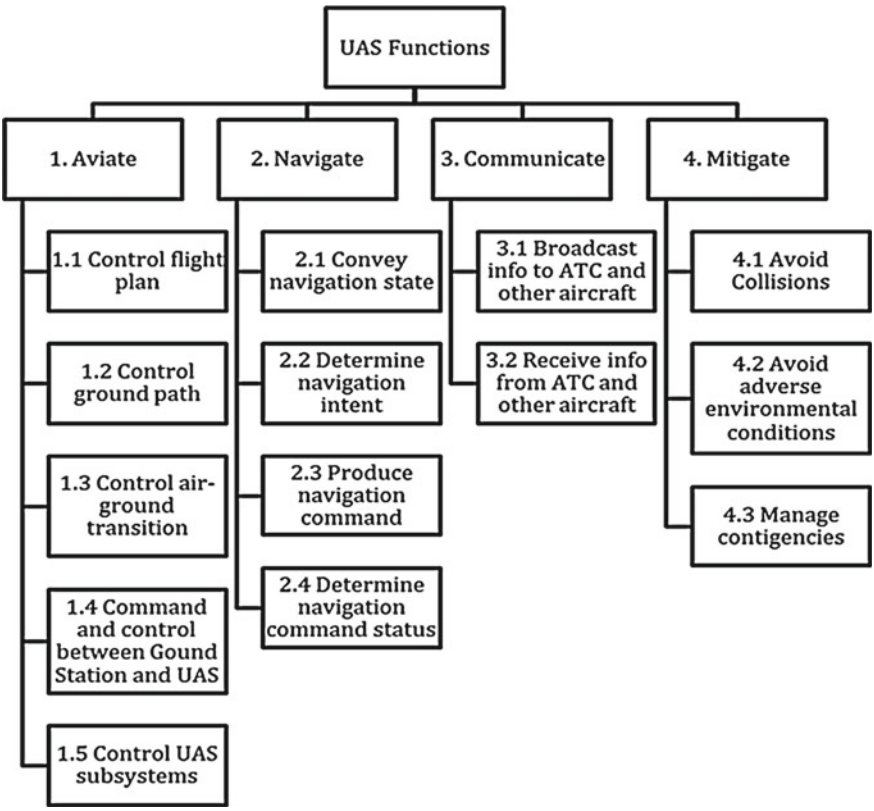


Fig. 14.4 Top-level of the functional decomposition (adapted from National Aeronautics and Space Administration 2007)

Table 14.3 Excerpt of FHA decomposition for "Control RPAS Subsystems" (adapted from National Aeronautics and Space Administration 2007)

Number	Function	Flight phase	Remark
1.5	Control UAS subsystems		
1.5.1	Control power subsystems	Enroute	The power subsystems are defined as the components that generate and distribute power such as electrical generator/distributions
	Failure condition	Operational consequence	Classification
1.5.1a	Total loss of function	Loss of the power subsystem may result in loss of control of ANTEX-X02 Extended UAS and possible out of control landing	Catastrophic
1.5.l.b	Misleading command	Misleading info to/from the power system may result in loss of ANTEX-X02 Extended UAS and possible out of control landing	Catastrophic

communication among the RPAS, the ATC and other aircraft. Mitigate involves traffic avoidance, ground objects avoidance, bad weather conditions avoidance or other types of environmental hazard impacts and helps solving mishaps (National Aeronautics and Space Administration 2007).

The RPAS level FHA and respective fault tree provide an initial group of failure conditions and related requirements to be taken into consideration at the system level. Afterwards, the conceptual design phase includes the architecture development based on FHA safety objectives. The Antex-X02 Extended RPAS functional architecture was outlined based on this model.

The full functional decomposition is relatively large, so as an example, for the function "Control RPAS Subsystems" (Table 14.3), a set of failure conditions and assumptions for the assessment were determined using the safety criteria indicated in Tables 14.1 and 14.2.

The results of the aircraft FHA are the inputs to the system FHA. The results of the system FHA are associated to the system FMEA inputs and to the PSSA process (an example is presented in Table 14.4). The decision to carry out a PSSA is dependent upon the design of the architecture, complexity, the severity of the failure condition (or failure effect), and type of function performed by the system being analysed.

Determining the likely failure modes of the functions of the Antex-X02 Extended systems allowed for the determination of the systems whose functions were critical to the safety of the RPAS operation. In this context, it was possible to develop and

Table 14.4 Excerpt of electrical power system FMEA

1—Antex-X02 EXT100_CIAFA				
1.2. Aircraft				
Function	Perform autonomous flight with imaging			
1.2.5 Electrical power system				
Function	Provide power to all UAS systems			
Failure mode	Loss of electrical energy (Electrical power system)			
Consequences	Systems does not function loss UAS			
Si	5			
Causes	Failure in energy supply	Breaking drive belt that connects the shaft of the propulsion system engine to the alternator	Alternator brushes breakdown	Electrical breakdown in the alternator management board
Causes mitigation	Perform readiness pre-	Completion of the preventive maintenance program		

Table 14.5 Excerpt of electrical power system list of mitigation processes

Item	Electrical power system	Electrical management board
Failure modes	Power failure (electrical power system)	Failure power supplying
Failure Prob.	Level B extremely remote	Level C remote
Clas. Severity	Category I—Catastrophic	Category I—Catastrophic
Function	Supply power to all platform systems	Manage the power supply
Mitigation actions	The operator turns the backup generator, and activates the alternative circuit and connects the support equipment	The operator transfers the platform control to the safety pilot so that it move on to manual mode. The safety pilot glides the platform for the nearest safe location. The operator informs the occurrence to ATC, for them to proceed to the diversion of traffic in the area
Mitigation process	Perform functional tests. Perform conditional maintenance program	Perform functional tests. Perform conditional maintenance program
Hazard	A	A

implement strategies, early in the design phase, to mitigate these vulnerabilities, namely the introduction of redundant systems and alert and monitoring systems.

For each failure mode classified as catastrophic, hazard and major, mitigation actions (performed during flight) and mitigation procedures (on the ground) were established, in order to achieve an acceptable level of safety, as presented in Table 14.5. In this case, the mitigation process has set up activities for both the design and the operator entities.

The results of the safety assessment have demonstrated that the Antex-X02 Extended RPAS was compliant with the minimum level of safety defined in STANAG 4703 for segregated air space missions, which has contributed, among other body of evidences, to obtain the permit to fly.

Future research will be focused on the study of the human error in unmanned aircraft systems, in order to integrate other safety issues.

14.6 Conclusions

The airspace rules for RPAS are the same as the ones applicable to manned aviation, with a specific mean of applicability and proportionality. A RPAS can be integrated in the correspondent airspace class, depending on its purpose and demonstrated level safety achieved. The main goal of a RPAS' designer is to develop a RPAS able of complying with the applicable rules of a particular airspace class, adequate for the target intent of use and missions. The RPAS operator is, ultimately, responsible for its correct use.

The AAN defines the minimum requirements to ensure that the RPAS operations are conducted within an acceptable level of safety. One of the requirements is the submission of a safety case with its substation. A manned widespread safety assessment methodology (SAE ARP 4761), according to STANAG 4671, 4702, and 4703, is considered as an acceptable means of compliance for such substation. The SAE ARP 4761 safety system concept comprises the definition of RPAS safety objectives at FHA and their validation and verification in PSSA and SSA respectively.

The application of this methodology to Antex-X02 Extended RPAS has set forth some failure modes which safety level were considered not acceptable. A deeper analysis, including the implementation of mitigation actions, has demonstrated that Antex-X02 Extended RPAS was compliant with a minimum safety level, specified in STANAG 4703, to fly in segregated air space. Afterwards, a permit to fly was issued by the AAN based on the safety assessment process, among other evidences (e.g. maintenance program, safety checklist).

Acknowledgements The authors would like to thank the CISDI research centre for the opportunity to present this paper. In addition, the authors gratefully acknowledge the CIAFA research centre for the possibility to present the safety assessment case study applied to the Antex-X02 Extended UAV platform. Such platform was developed by a joint team of the Portuguese Air Force Academy and Oporto University, under the PITVANT research project, funded by Portuguese Ministry of Defence.

References

Autoridade Aeronáutica Nacional (2013) Emissão de Licenças Especiais de Aeronavegabilidade, Circular No. 1/13

Center for Advanced Aviation System Development (2012) Unmanned aircraft systems: airspace integration. MITRE

Civil Aviation Authority (2002) CAP 715: an introduction to aircraft maintenance engineering human factors for JAR 66. https://publicapps.caa.co.uk/docs/33/CAP715.PDF. Accessed 28 Apr 2015

Cortright D, Fairhurst R, Wall K (2015) Drones the future of armed conflict: ethical, legal, and strategic implications. University of Chicago Press, Chicago

DeGarmo M (2004) Issues concerning integration of unmanned aerial vehicles in civil airspace. https://www.mitre.org/sites/default/files/pdf/04_1232.pdf. Accessed 28 Apr 2015

Department of Defense United States of America (2012) MIL-STD-882E. System Safety, Standard Practice

Dhillon B (2009) Human reliability, error, and human factors in engineering maintenance. CRC Press, New York

Ebeling C (2005) An introduction to reliability and maintainability engineering. Waveland Press Inc., USA

European Aviation Safety Agency (2005) A-NPA, No. 16/2005, policy for unmanned aerial vehicle (UAV) certification. https://easa.europa.eu/system/files/dfu/NPA_16_2005.pdf. Accessed 28 Apr 2015

European Aviation Safety Agency (2015) Concept of operations for drones—a risk based approach to regulation of unmanned aircraft. https://www.easa.europa.eu/system/files/dfu/204696_EASA_concept_drone_brochure_web.pdf. Accessed 2 Apr 2015

European Commission (2014) Study on privacy, data protection and ethical risks in civil remotely piloted aircraft systems operations. Final Report, Brussels, Belgium

Evans A, Nicholson M (2007) Safety assessment and certification for UAS. In: Proceedings of 22nd Bristol UAV systems conference, Bristol, UK

International Civil Aviation Organization (2011) Circular 328, Unmanned Aircraft Systems (UAS). Quebec, Canada

International Civil Aviation Organization (2013a) SAFETY—2014-2016 global aviation safety plan. Document 10004, Montréal, Canada

International Civil Aviation Organization (2013b) Safety Management Manual (SMM), Doc 9859 AN/474. http://www.icao.int/safety/SafetyManagement/Documents/Doc.9859.3rd. Accessed 28 Apr 2015

Jones M, Melding D (2014) FAA system safety—blending SAE ARP4761 and MIL-STD-882E processes. In: Proceedings of international system safety society conference. St, Louis, USA

National Aeronautics and Space Administration (2007) Preliminary considerations for classifying hazards of unmanned aircraft systems

NATO Standardization Agency (2009) STANAG 4671—unmanned aerial vehicle systems airworthiness requirements. Edition 1.0

NATO Standardization Agency (2014a) STANAG 4702—rotary wing unmanned aircraft systems airworthiness requirements. Edition 1.0

NATO Standardization Agency (2014b) STANAG 4703—light unmanned aircraft systems airworthiness requirements. Edition 1.0

Pumfrey D, Dawkins S, McDermid JA, Murdoch J, Kelly T (1999) Issues in the conduct of Preliminary System Safety Analysis (PSSA). In: Proceedings of 17th international system safety society conference, Florida, USA

Reason JT (1997) Managing the risks of organizational accidents. Ashgate Publishing Limited, Aldershot

Society of Automotive Engineers (1996) SAE ARP 4761—guidelines and methods for conducting the safety assessment process on civil airborne systems and equipment

Chapter 15
Robots and the Military: A Strategic View

João Vieira Borges

Abstract This paper views the theme of robotics in the military domain from a strategic perspective, bearing in mind the new paradigm of security and defense, where robots will have an increasing intervention. Considering the trilogy that strategy comprehends—goals, means and threats, three fundamental topics are approached: (i) the need to work at political, strategical, operational and tactical levels (ii) the role of robots in the new security and defense environment (iii) the importance of incorporating robots in the military formation. As a conclusion, this paper also highlights the importance of introducing the major issues associated with robots, from artificial intelligence to robot ethics, in the curricula, research and training carried out at military schools with the purpose of preparing military commanders for a future where robots will most likely have a prominent role.

Keywords Robots · Military strategy · Threats · Security and defense · Education

15.1 Introduction

As a strategy professor I usually think, investigate, write and talk about strategy as a science and an art, but particularly about the national security strategy of the states.

National security strategy deals with the potential and the vulnerabilities of states, in order to ensure their independence and sovereignty in ever-evolving internal and external environments where new threats and risks emerge.

The present world, with its new relationship between economics and politics, a change of global agenda, a growing technological capacity, actors other than states with a high capacity for destruction and, consequently, different threats and risks at global level requires a reevaluation of strategy. It is indeed a new paradigm, in which the means, goals and threats-the strategy's trilogy-are very different from what they were a few years ago.

J.V. Borges (✉)
Portuguese Military Academy, Lisboa, Portugal
e-mail: borges.jjbv@mail.exercito.pt; joaovieiraborges@gmail.com

M.I. Aldinhas Ferreira et al. (eds.), *A World with Robots*,
Intelligent Systems, Control and Automation: Science and Engineering 84,
DOI 10.1007/978-3-319-46667-5_15

To face those threats and risks with our grounding values (freedom, democracy, rule of law, human rights) in a different political and military environment, we must have different "soldiers", different equipment and weapons, and different tactics and doctrines.

In the concept of the new "soldiers" and weapons systems, the robots (as a device or group of electro mechanical or biomechanical devices able to perform autonomous pre-programmed work)[1] should be included without forgetting the main ethical issues associated with this new reality.

In the next lines we will cover three major strategic issues: (i) the need to work the four levels of approach-political, strategical, operational and tactical (ii) robots in the new security and defense environment (iii) the importance of working robots in the military professional qualification.

Our final remarks address the importance of introducing the topic of technology warfare its role and ethical constraints in the curricula of the military academies responsible for the professional education and training of the cadets that will be the future officers and military leaders.

15.2 The Four Levels of Approach

Singer (2009) organized his book "Wired for War" in two chapters: "The Change we are creating" and "What Change is created for us". He talks about robots in war, ethics, law, the demographics of conflict, command, control and autonomy. These are important issues, but a strategic vision on robots should start by an integrated approach comprehending four different levels: political, strategical, operational and tactical.

At political level, it is important to talk about the relationship of robots with values like freedom, democracy, rule of law and human rights, national interests and global/regional/national strategies.

The context of values is the most complex, even though the majority of the states shares the concepts of freedom, democracy, rule of law and human rights. Interests are directly related to the availability of means and to the constraints to achieve them, which suggests that the use of robots will be restricted to a few actors, who will be in the vanguard and thus in strategic advantage. In the context of global/regional/national strategies, the differences are self evident. For instance, at national level, it is completely different to talk about Global States as the USA (with a capacity for global intervention at the political, economic, diplomatic and military level) and states such as Portugal.

The Global States have the capacity for planning, researching, producing, testing, financing and using technology in concerted and integrated actions. On the other hand, the majority of countries, like Portugal, only partially produce, and most often acquires technologies and adjust the tactics.

[1] UAVs (drones) are not yet classified as robots according to NATO (2015).

At the strategic level we must evaluate the benefits and drawbacks of the use of robots to increase the potential or reduce the vulnerabilities of states. We must justify and integrate the use of robots in different situations.

The operational level regards a more specific area. When we discuss the military area for example, we have to think whether to include (or not) a battalion of predators, 50 % of armed robotic vehicles, and so on. At this level we must consider the battle and not the combat, adopting a joint and multinational approach.

Finally, at the tactical level (of the combat) we talk about doctrine sustaining the use of military robots in different military applications. Robots know no fear, do not get tired, do not eat, do not close their eyes, do not talk to their friends, respect all the orders, but we must integrate them in the global doctrine, in which man continues to be the center of the combat.

Usually studies on the use of robots in the military domain focus on the role of artificial intelligence and ethical issues arising from their use. To reflect on the usage of robots an integrated approach is fundamental, because there are significant and consequent relations (not only hierarchical) between the different components of a very complex framework.

15.3 Robots in the New Security and Defense Environment

Today, we have a more complex world. In addition to the states, we have a plurality of actors with greater weight in the international political system, including international organizations, non-state organizations and even citizens.

Moreover, the progressive acceleration of globalization in its different dimensions (political, economic, social, etc.), is directly related to the issues of security and welfare in the world, especially because of the crescendo of global threats such as transnational terrorism, transnational organized crime and, among others, weapons of mass destruction.

Our world, in this first quarter of the twenty-first century, is characterized by unpredictability and volatility, and continues to have in the US the "global power" with a global intervention capacity. In geostrategic terms, there are two other actors with significant strategic importance, namely Russia and China, which are the main opponents to the US hegemony, at a time when Europe and countries such as the UK, France and Germany are wrapped with a continuity of character, in a financial and values' crisis. This "trilogy" has allowed the US investment in the relocation of world power from the center of Europe and the Atlantic to Asia and the Pacific.

However will the "trilogy" USA-Russia-China dominate the new world changes? Certainly not, as demonstrated by other non-state actors, who regularly threaten the security and stability of the international political system. In fact, this is characterized, presently, by the following features:

- the existence of non-state legal entities, particularly unmasked terrorists, as those from al-Qaeda to ISIS, that fighting without rules or ethical concerns aim at the destruction of Western patterns of social life and civilizational framework.
- the globalization of threats and risks. Globalization brought to the world the dissemination of transnational threats and risks such as terrorism, weapons of mass destruction, transnational organized crime, piracy, pandemics, fragile states, civil wars, disputes over scarce natural resources, climate change and cybercrime. Global threats today are closer to citizens, as they are less territorial, more demilitarized and even more difficult to identify and characterize. Global threats, particularly transnational terrorism, have marked the agenda of the international political system, especially after 11 September 2001.
- the era of information and knowledge. Nowadays we have an extremely high technological capacity which can either save lives or create serious security problems.
- the increasing importance of the economic context. Reality shows how the economy increasingly determines policy options and often affects the interests and the values of states, with visible consequences not only for the state but also for citizens.
- the clear location of the center of gravity of world power, from the Atlantic to the Pacific, not only determined by the US interest but also by the economy.
- the demographic clash between the North "Rich but Old" and the South "Poor but Young", in a world that, is, nowadays, more urban, more aged, less secure, and has less values and greater social inequalities.
- a weakened state less able to adequately protect its people against threats and risks, generating a sense of general insecurity.
- the "power of circumstance". The globalized world of today, holder of so much information, is marked daily by the "power of circumstance", taken by the images of the Paris or Brussels attacks, the massacres perpetrated by ISIS on wasteland and the shocking images of refugees disembarking in lands of Europe, leaving aside the war in Ukraine, the war between the Koreas, or the approach between the US and Cuba.
- the drivers of change. For NATO (and especially for the Allied Command Transformation), drivers of change are the future of the instruments, which need to be identified and characterized as assumptions of scenario construction. In the most recent NATO strategic concept, the drivers of change identified and studied were demographic change, globalization, limited resources, climate change and the use of new technologies.

Faced with this synthetic characterization of the international political system, it is understandable that it is not only time to return to political ideas, to invest in technology and knowledge, but also to defend values such as freedom, democracy, rule of law and human rights, even with the individual and collective sacrifice, in order to have better future.

In this global environment, we have difficulties to separate the boundaries of security and defense. With the global threats as international terrorism or international organized crime, we never know the space of intervention of the Armed Forces and

the national polices and gendarmeries. The missions of the Armed Forces and Police are complementary in many cases, leading to greater difficulties in identifying their systems of forces and weapons systems.

To protect the population and property, in this urban world, the police already uses robots to protect people whose lives are at stake in missions, such as remote controlled devices[2] designed for internal security applications. The Armed Forces also use robots in demining mission's minefields or remote controlled devices.

But security and defense must work together (in an integrated way) and coordinate the use of robots, despite the differences between actors. This is why militaries and polices need to work in a more integrated and coordinated way to be more effective and efficient, as summarized below:

- In the case of polices, robots are needed to save and preserve human life on both sides: The police and the people. As the military, they need surveillance, sniper detection, neutralizing explosive devices or improvised Explosive Devices (IEDs).
- In the case of Armed Forces, robots are needed to protect soldiers and in specific cases to destroy the enemy. Robotic systems are now widely present in the modern battlefield,[3] providing intelligence gathering, surveillance, reconnaissance, and target acquisition, designation and engagement capabilities Arkin (2013). Multiple potential benefits of intelligent war machines have already been declared by the military, including a reduction in friendly casualties, force multiplication, expansion of the battlefield; extension of the warfighter's reach, the ability to respond faster given the pressure of an ever increasing battlefield tempo, and greater precision due to persistent stare[4] Arkin (2013). According to some authors, such as Singer (2009), the robots will form the armies of the future and, depending on its nature, will be configured to do the "dirty work". In "Wired for War", Singer argues that just like the tank and the submarine in World War I, or the atomic bomb on World War II, the military robot will be seen as the greatest war technological innovation.

In both cases it is time to think about ethics. To take someone's life, even in a war situation, is something that has to be judged by society and international community, especially when it comes to robots programmed by man and delegated intervention capacity built by artificial intelligence. It could be unethical and would

[2]IED—Improved explosive devices. In Iraq it became the leading cause of casualties among US troops as well as Iraqi people. In response, the Pentagon soon was spending more than 6.1 billion dollars to counter IEDs in Iraq. The merit goes to the EOD (Explosive Ordnance Disposal) team who saved many lives (with the help of robots).

[3]The Big Dog is a military robot project that is being developed by Boston Dynamics, becoming a kind of "mule-without-head" load. This robot can carry up to 180 kg in its structure; in addition, it reacts to movements around it, not falling when pushed or slipping, managing to walk on any terrain (including ice). It is already used by the military in some places in the world to transport equipment and in the future it will be marketed.

[4]In October 2007, a robot (antiaircraft equipment Oerlikon GDF-005) killed 9 people and injured 14 others seriously in a test performed on a military base in South Africa (Lahotlha). The cause was software failures.

lead to violations of human-rights laws, as well as, international laws governing combat. The US has internal directives circumscribing the usage of autonomous and semi-autonomous weapons. Arkin (2013) proposed software architecture for introducing ethics into autonomous weapons systems, but he also acknowledges that in the future, autonomous robots may be able to perform better than humans under battlefield conditions Arkin (2009).

15.4 Robots and the Military Professional Education

When we talk about robots, we always mentally positioned ourselves in the future. But the future is built at the present, consequently we think that the use of robotic technology in the military domains should be studied and investigated in military schools that prepare the military official (leaders) of the future.

When facing technological change, some authors predict that robots may begin to replace humans in many areas with intensive work, including in the security and defense area. Asimov (1950) even created the Three Laws of Robotics, in an attempt to rule the coexistence of human beings and intelligent machines:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm;
2. A robot must obey the orders received by human beings except where such orders would conflict with the First Law;
3. A robot can protect its own existence as long as such protection does not conflict with the First or Second Laws.

The introduction of the course of "robot ethics" or "artificial intelligence" in the master's degree of military sciences is increasingly important. We need to improve the education and prepare the young future military leaders for the new world, the new threats and risks, and the new military environment.

Nowadays the cadets and future military officials in West Point (USA), in Sandhurst (UK), in Saint Cyr (France), in Zaragoza (Spain) or in Lisbon (Portugal), study mathematics, chemistry, physics, management, leadership, decision theories, sociology, psychology, international relations, and strategy, besides the military and behavioral training. However, they need to understand the use of technology as a new combat tool, as it is the case of robots, and relate them with ethics and the rules of engagement. In the light of the growing usage of robots, their programs should also include the courses on robot ethics, artificial intelligence or even deployment of robots on the field.

This holistic approach to different scientific areas is critical in higher military education in order to prepare military officers for the future. This multidisciplinary approach is certainly particularly suited to the new paradigm of security and defense in general.

15.5 Final Remarks

In the new world marked by global threats, which are determining the new paradigm of security and defense, the role of technology in asymmetric warfare will be increasingly important.

For Singer (2009), by 2025, the robot industry might rival the automobile and computer industries in both dollars and jobs. In this sense the use of robots in war will materialize technological developments and will make a real revolution in military affairs.

Even though this possibility is still a few years ahead, it is fundamental to study and reflect on the ethical issues that emerge from this particular domain, in particular those relative to their deployment as combatants taking the lives of enemies and especially of innocent noncombatants.

The approaches to the deployment of robots in military contexts should not be only focused on the technical or the psychological perspectives. The most adequate vision should also include political, strategical, operational and tactical perspectives, always taking into account the values, the interests, and strengthening the potential and reducing the vulnerabilities of states.

The growing usage of robots by the police and armed forces also implies a better and increasing coordination between the different actors within the new security and defense paradigm.

Teaching and training topics relative to the future of war are fundamental at military academies responsible for the professional education and qualification of future military officers. Therefore, it is important to create new courses that equate all the variants present in the present complex civilizational framework and that take into account all the research and thought developed in what relates the ethical and societal problems involved in technological warfare. Among the subjects and research areas to be included should not only be artificial intelligence, cyber defense and information, but also ethics, namely robotethics, because despite all technological development the centre of decision and action will always be human beings.

References

Arkin R (2009) Governing Lethal behavior in autonomous robots. CRC Press, New York

Arkin R (2013) Lethal autonomous systems and the plight of the non-combatant. AISB Q (137). https://smartech.gatech.edu/bitstream/handle/1853/50079/aisbq-137.pdf?sequence=1. Accessed 25 Mar 2016

Asimov I (1950) Robot. Gnome Press, New York

NATO (2015) NATO/multinational joint intelligence. A feasibility study, surveillance and reconnaissance unit

Singer P (2009) Wired for war: the robotics revolution and conflict in the 21st century. The Penguin Press, New York

Chapter 16
The Domestic Robot: Ethical and Technical Concerns

Rodolphe Gelin

Abstract In the present paper we want to highlight the importance for the social/service robotics designer of being aware of the potential ethical and safety issues that may arise from the development of humanoid robots functioning as companions. After a short description of a possible use case, dedicated to the assistance of an elderly person, we identify the main concerns from the safety and ethical point of views and propose ways on how to prevent risks.

Keywords Robot companions · Essential design features · Ethics · Elderly rights

16.1 Introduction

Having robots populating our daily life in the close future raises several questions about the coexistence between man and machine. Being robotic developers and producers we want to share in this paper not only the way we envisage life with robots but mainly how SBR views and addresses the possible safety and ethical concerns that may arise from a close human robot interaction.

Following a full day in the life of a senior assisted by a humanoid robot allows us to identify (i) the important features a robot should be equipped with, in order to help lonely elders to be autonomous, in spite of their light cognitive or physical impairments (ii) how to effectively look after them in the context of limited contacts with other caregivers, guaranteeing essential human rights as their safety and privacy.

This use case has been studied by SBR and its partners within the framework of a number of collaborative projects such as ANR Riddle (2015), FUI Romeo (2015), PSPC Romeo (2015).

Robotics, namely the production of social/service robots has been acknowledged as a possible solution to face the crucial problem of an ageing society, but we have to make sure that solving partially this problem will not generate other kind of problems

R. Gelin (✉)
SoftBank Robotics, Paris, France
e-mail: rgelin@aldebaran.com

M.I. Aldinhas Ferreira et al. (eds.), *A World with Robots*,
Intelligent Systems, Control and Automation: Science and Engineering 84,
DOI 10.1007/978-3-319-46667-5_16

involving legal, societal or ethical aspects. As robot manufacturer, SBR is aware of the complexity of the task. In fact, the definition of a roadmap where the ethical and societal impacts of robotic technology are safeguarded can only be achieved with the cooperative work of the robotic community and the rest of society. The present paper aims to be a contribution to this.

16.2 Mrs. Smith and Her Robot—Describing a Daily Routine

Mrs. Smith is 85 years old. She lives alone. She is in rather good shape for her age but she has slight memory troubles. Her movements have become slower and more painful and she has some trouble maintaining her balance.

When Mrs. Smith wakes up, the robot, which was recharging its battery outside the room, detects her activity and comes to the door-way of the room to greet her. Detecting her mood, the robot knows if it should stay outside the room and wait for her or if it can come in to have a chat about how her night was. As Mrs. Smith seems to want to stay longer in bed the robot keeps chatting with her.

When Mrs. Smith wants to get up, the robot helps her to maintain her balance as she gets out of bed and then steps out of the way to let her walk out of the bedroom by herself. In the kitchen, Mrs. Smith fixes herself a coffee and takes the newspaper. The robot understands she will need her reading glasses that she forgot in her bedroom. It goes there and brings the glasses back to her (Fig. 16.1).

When her breakfast is finished, the robot proposes that Mrs. Smith does some exercises which she accepts. Based on the recommendation of her physiotherapist, the robot demonstrates some soft and slow stretching motions that Mrs. Smith can replicate. The robot broadcasts steady paced music to accompany the movements. It detects that Mrs. Smith is in a good shape today. Her motions are pretty good and she looks happy.

Afterwards, Mrs. Smith goes to the bathroom. For privacy reasons, the robot stays outside but close enough to the door to detect potential unusual noises. The robot notices that Mrs. Smith stays longer than usual in the bathroom. But it recognizes usual noises and just waits. Two days ago, Mrs. Smith stayed unusually longer and silent in the bathroom. The robot asked her through the door if everything was fine. She said yes but she was exhausted by the gym and needed to rest a little bit before finishing her make up.

For the rest of the morning, Mrs. Smith reads her book and calls her friends to chat and to confirm their bridge game at about 5 pm. The robot does not do anything but proposes to turn on the light because the weather is cloudy and it will be difficult to read with only the day light. At 11am, the door bells rings. It is the woman who assists Mrs. Smith with her shopping and groceries, once a week. The robot asks Mrs. Smith if she wants it to open the door or if she wants to go. Mrs. Smith, still full of energy this morning, goes by herself. The lady stays half an hour to unpack, put together the shopping and chat with Mrs. Smith.

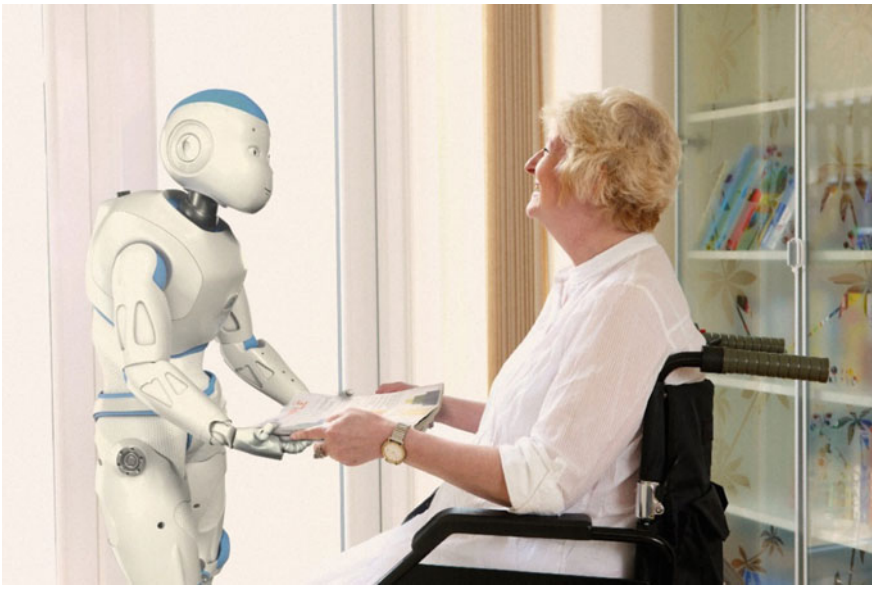


Fig. 16.1 Mrs. Smith and her Robot

During lunch, Mrs. Smith is in such a good mood that she wants to drink a glass of wine to celebrate (her father was French). The robot mentions to her that it is not recommended to drink too much if she wants to win her bridge game this afternoon. Mrs. Smith answers that just one glass should not diminish her capacity to make a grand slam to her opponents. But after the lunch, Mrs. Smith is exhausted by this long morning and the red wine and she goes for a little nap.

During the nap, the robot stays outside the bedroom. Generally, Mrs. Smith's nap is less than an hour long. After an hour and a half, Mrs. Smith does not wake up. The robot goes into the room and calls her softly but she still does not react. As planned in such situation, the robot calls the assistance center. The remote operator takes access of the information gathered by the robot. He takes the control of the robot. Very gently, using his force feedback master arm to control the hand of the robot, he touches the shoulder of Mrs. Smith. Mrs. Smith wakes up. She is fine, she has just spent too much energy in the morning. The operator recommends her to have a little visit outside to get some fresh air. Mrs. Smith says it is a good idea, puts her coat on and goes around.

At 3:45, Mrs. Smith is back into the house. She sits for a while in front of the TV to regain some energy from her walk before preparing tea and cakes for her friends. At 5, they arrive. They start chat-ting all together. The robot stays away to leave them alone but positions itself in a way if Mrs. Smith looks at it, he will be able to assist her. After a while, Mrs. Smith calls the robot and asks him to bring the cards. During the game, the robot takes pictures.

When all her guests are gone, Mrs. Smith put things away with the robot and eats a quick dinner. As she is too tired to go into her bed-room, she sits in front of the TV. The robot proposes a TV program and gives her glasses. She reads a summary of the program and turns on the TV. But after 5 min, she is already sleeping. The robot notices it. When it detects that Mrs. Smith is moving a little bit, it calls her and recommends her to go to bed. She agrees. She stands up with the help of the robot and walks slowly to the bath-room then to her room. The robot wishes her a good night and leaves the room saying it will be around if she needs it.

16.3 The Robot's Social Behavior: From Proactive to Intrusive

One of the major advantages of the robot, compared to other digital devices, is its active sensing capabilities. It can move around to look for useful information, it can stay close to Mrs. Smith to understand what she is doing and propose the right service at the right moment, e.g., suggest to turn the light on when it is dark.

Within the framework of the aforementioned projects, when inquired about their expectations relatively to a domestic robot's behaviour, some elderly referred that they would not like to have a proactive robot that starts moving around without any understandable reason. If the robot is expected to be proactive, it would have to explain its user what it is doing and the reason why it is doing it. The other concern relative to the robot's proactivity is the discomfort caused by the feeling of being permanently observed by a machine. That is an effect similar to that caused by video surveillance systems where the camera is clearly located. However if the robot is not is the same room, it can't see its user. If Mrs. Smith wants to be alone and to be certain of her privacy, she can always ask the robot to stay outside the room or at a short distance to be available when needed. However, this feeling of intrusiveness has rational reasons- if the robot stores what it perceives, it can broadcast it and show it to other people, in real time or later. Mrs. Smith is in fact virtually not alone with the robot. In order to dissipate her aprehension she would need to know that the robot does not store and does not broadcast what it sees and hears unless she allows it to happen. It is possible to imagine a special signal being emitted by the robot (like a blinking LED) indicating when this storing is active and when it isn't.

Another sensible issue is the elder's eventual monitoring by an external entity. For instance, as mentioned above, during Mrs. Smith's longer nap, the robot calls a remote operators intervention in order to check if everything is all right with the user. This operator follows the health status of Mrs Smith and is only in (virtual) contact with her when the robot has evaluated that the situation requires the operator's intervention. However in terms of privacy and personal security this potentially involves a certain amount of risk. We can consider that the operator is a professional and does not care of what Mrs. Smith is doing at home, but if an ill-intentioned person takes control of the robot, this becomes very dangerous situation. Not only for the privacy of

Mrs. Smith but also for her safety. A thief can look where Mrs. Smith hides her jewelry. It can control the robot to open the door to a stranger when Mrs. Smith is out for a walk. The robot manufacturer has to provide all that is necessary to prevent this kind of risks, installing the necessary protection against "spyware" and encrypted communications between the robot and the outside world. Robotics should be fed by all the work created for computer safety.

16.4 Physically Interacting with the Real World

The unique characteristic of the robot, compared to other digital devices is its capacity to move around by itself and to move other objects. This gives the robot the possibility of staying close to its user, of fetching objects but also, in the near future, of physically assisting someone like Mrs. Smith. However, several scientific and technological challenges still need to be tackled before these features are available on a domestic robot.

To be efficient in our world, the robot has to be tall enough to grasp objects on a table, to manipulate a door handle and to interact in a natural way with human beings (if the robot is too small, Mrs. Smith has to lean forward to talk to it). The "correct" estimated size for a humanoid robot is between 0.9 and 1.20 m. Being tall, the robot has to be strong enough to carry its own weight (probably around 20 kg). Furthermore, if it has to be able to assist Mrs. Smith, for instance when she moves, it has to be able to sustain part of her weight. We gave up the idea that the robot could carry Mrs. Smith. Even if this kind of robots exist Mukai (2010), they have to be very powerful and bulky. They are not adapted yet for domestic applications. But even if our robot is not able to carry 70 kg, it has to be powerful enough to provide 10 kg worth support to Mrs. Smith. This makes the robot powerful enough to be potentially dangerous. This is a dilemma for the designer: to be really useful the robot becomes potentially dangerous. The 58 cm high robot NAO Gouaillier (2009) is not dangerous but it is not a realistic solution for most of the physical interactions expected from a humanoid robot. The Atlas robot Feng (2015) is very impressive and very dynamic. It is able to carry 10 kg loads but it is also frightening and probably dangerous.

Technical solutions exist to prevent the risk of unexpected and dangerous physical interactions between the robot and its user. At the base of them is the ability for the robot to detect the presence of a human being (before the contact) and detect the contact when it occurs. Unlike industrial robots, the service robots may interact physically with human beings. Either by using tactile sensors or by man-aging a dynamic model that computes the required force for its cur-rent task Albu-Schaeffer et al. (2008), the robot has to be able to estimate if the torque it applies is normal or not and stop the motion if the torque is not normal. It has to be aware of its own strength. This is part of the awareness robots should be endowed with by the designer.

An important ethical issue that emerges within the assistive framework is: how assistive the robot should be? Some of the elders inquired have expressed their concern that an excessively prestative robot could lead seniors to rely on robots for

everything, even for the tasks they could perform well causing them to lose more quickly their remaining physical abilities and, in some way even infantilize them. To prevent this risk, the robot has to take in consideration the current ability its user. If she is still strong enough to pour water in her glass from the bottle by herself, when she asks for water, the robot has just to bring the bottle to her, and then she can pour the water by herself. This perspective taking Pandey (2015) is also a possible solution to maintain the robot at the good position within its relationship with the humans.

16.5 Feelings and Emotions

When SBR presented NAO, in 2006, it came up as a major novelty in the domain of humanoid robotics: NAO was a cute humanoid robot. It had been designed to be a robot companion meant to be accepted as a nice little fellow. Before NAO, the existing humanoid robots were impressive, similar to astronauts Doan (2015) or like science fiction warriors Kaneko (2015) but none of them was cute or aimed to enhance any affective response from users. In Japan, the robotic baby seal Paro Bemelmans (2015) dedicated to the well-being of institutionalized elderly people had enhanced the affective engagement of its users, but it was not a humanoid. To overcome the prior stereotypes people generally have towards humanoid robots, a nice design/appearance was necessary to make it accepted. Beyond the appearance of the robot, its behavior is important. By its gestures, its voice, the way it speaks, the companion robot becomes more than a simple object. It interacts with the user using the same communication codes that humans do. Yet, the expression of emotions is an important part of the communication between humans. So the robot also has to be able to recognize human emotional states and also be able to express basic emotions itself. NAO naturally achieves this with its appearance but developers even improved this ability by giving it the ability to recognize the emotion of its user and to generate the adequate expressive responses.

When the robot interacts with Mrs. Smith in the morning, its embedded software tries to detect, through audio and video streams, clues that indicate if she is happy, sad, grumpy or tired. Thanks to a learning process, the robot can classify the psychological status of Mrs. Smith Devillers (2015) and adapt its behavior to her the emotional mood. The users emotions become one more input in the decision process of the embedded software. The robot no longer reacts just in a machine-like way. Taking into account the emotional status of Mrs. Smith, the robot does not just perform functionally but adequates its response to a specific affective context, being even capable of giving the user the impression of caring and creating this way an emphatic relationship. This is a risk of using feeling as a parameter in the human-robot interaction. However, in Tisseron (2015), the author indicates that soldiers can develop a kind of affection for the demining robots that saved their lives, even if these robots are simple remotely controlled gray boxes on tracks. In that particular case, the manufacturer of the robot cant be considered responsible for the excessive attach-

ment. But when the developer deliberately uses the emotion as part of the human robot interaction, he has some responsibilities, namely the responsibility of warning the user to the necessity of interacting with other human beings, the necessity of keeping their net of social/affective links with other human beings. By checking the amount of time spent with the robot and the kind of activities carried out by the user, the robot should evaluate the quality of his/her social life. In the same way that a navigation system in car warns the driver when he has driven more than 2 hours without a rest, the robot should say: "We have spent two days together and you did not talk to anyone else during these two days. Why dont we call your son or one of your friends?"

By the way it talks, it moves and the kind of gestures it performs, the robot can mimic emotional states. Expressing emotions, the robot simulates having emotions. If the robot manufacturer pretends that his robot has emotions there is an amount of deception involved in this. This is probably the most crucial ethical aspect of the companion robot. But in fact the targeted goal is to replicate what psychologists call affective empathy Tisseron (2015). This is the first level of empathy, developed by very young infants (one and a half years old): the child detects your emotional state and replicates it. This simple mechanism makes you understand that the child perceives your happiness or your sadness. In the same way, a robot looking sad when users look sad informs that it detected that they are sad. This is a much more intuitive way to transmit this information than saying, with an inexpressive and metallic voice: "I see you are sad". We have to admit that the borderline is not clear between mimicking emotions and pretending to have emotions. Humans are fast to anthropomorphize objects and adding human features to them. Denis Vidal (2012) proposes that the user makes an anthropological pact with the robot: "I know you are a machine but I will act as you are a human". This requires a detachment that can be reinforced by the appearance of the robot. The robot should keep looking like a robot and should not look like a real human. That is why the work of Ishiguro (2015) and Hanson Sandry (2015), as beautiful they can be, are probably a risky direction for a sound relationship between human and humanoid robots. Another way to keep this distance is through education: the more the user knows about how the robot works, the more he can relativize his feelings towards the robot.

The last ethical aspect we want to address is the global one "Is it ethical to leave my grandmother with a robot?" It is probably not an ideal solution but today our society does not provide many other solutions to maintain elderly people at home, in safe conditions. Considering the size of the apartments in modern cities and the western way of life, there is no realistic way to have someone taking care, 24 hours a day, of someone who has lost their autonomy. When it happens, it is generally like a hell for the caregiver who doesn't have any time to rest. If the robot can give few hours of relief to the family, the benefit is unquestioned. Maybe someday, we can change our way of life to spend more time with our elderly, but in the meantime, robots and other technological devices are good solutions to offer them safety and dignity.

16.6 Accountability

The question of accountability is always mentioned within the ethics framework. But responsibility is probably more a legal than an ethical issue. If the robot hurts someone, who is responsible?

Because the robot is a device, as smart and as sophisticated as it can be, the manufacturer is responsible for its working correctly. If the robot hits Mrs. Smith because its anti-collision sensor does not work, the manufacturer is responsible. It should have used more reliable sensors. The manufacturer is supposed to guarantee the proper performing of the basic functions during the normal use of the robot. But with complex humanoid robots, it becomes more and more difficult to determine what a 'normal use' of a robot is. The manufacturer will have probably to provide a long list of warnings to prevent any accident. The user guide of the robot will be a very large book that no one will read but that every robot user would be supposed to have read. Furthermore, considering that the robot is a platform on which applications, developed by a third part, can run, who is responsible in case of trouble: the developer of the application or the manufacturer of the robot? After an accident, experts will have to investigate what really happened to determine if the application gave a bad command to the basic functions of the robot or if the robot failed in performing the required function. It means that the robot should have a kind of black box, like airplanes, storing all the information that can be collected to trace what had happened on the robot. The law will probably have to impose the black box on robots.

In some cases, the user of the robot can also be held responsible. If Mrs. Smith asks her robot to pour hot tea on the head of Mrs. Martin because she cheated during the bridge game, Mrs. Smith will be responsible of the damages to Mrs. Martin. Isaac Asimov (1950) imagined three laws of robotics to prevent this type of problem. According to the first law, "the robot should not hurt a human being". The robot is then supposed to evaluate the consequences of its actions to avoid hurting a human being. One can imagine how complex it can be for a robot to evaluate the consequences of its actions. Considering that the pain can be physical but also psychological, how could a robot qualify as good or bad the consequences of its actions? It is sometimes difficult even for a human being to determine if any action is good or bad, how could a robot do it autonomously? Furthermore, if it would apply ethics rules to determine its actions, where should come these rules from? From the manufacturer or the application developer? How could the user be sure that he is sharing the same ethics standards than the manufacturer or the developer? Beside the huge technical handbook of the robot, another huge ethical handbook should be provided as well. Last, but not least, if we imagine that it will be possible to implement ethic rules in the 'brain' of the robot, it means that Mrs. Smith does not have to question herself about the morality of the task she orders to the robot. The robot will decide if the action is good or bad. Mrs. Smith does not have to think about it anymore. Delegating the moral evaluation of our acts to a machine is giving up an important part of our humanity. If we consider that the ethics applied by the robot

is the ethics of the developer, Mrs. Smith adopts forever, and without questions, the ethics of somebody else as her ethics. This is the end of her freedom to think. For all these reasons, the robot should not have its own ethics.

16.7 Conclusion

The presented scenario is not especially sophisticated but probably won't be totally available as a service provided by a commercial product in the next 5 or 10 years. It is not the purpose of this paper to detail the technical reasons that lead to this delay, but we draw the attention of the reader on this schedule because it means that engineers and researchers as well as lawyers and the society as a whole have time to work on ways to anticipate and mitigate risks. The community of robotics companies does not say there is no risk with robots but we say that it is part of our responsibility to evaluate them, to take them into account and to develop the necessary research to minimize them. Today a car manufacturer does not propose a super-fast car without super-powerful brakes (because it is logical and commercially clever) and without efficient airbags (because it is mandatory). The robot manufacturer will have the same constraints: economical (it is difficult to sell a robot that does not make its user happy) and legal (it has to be forbidden to sell a robot that is intrinsically dangerous).

The analogy with the cars can be extended to the question of accountability. The first person responsible in an accident is the driver (if he made a mistake), then the manufacturer (if a failure had provoked the accident). The boundary between cars and robots is be-coming fuzzier. The cars are already able to park by themselves, to decide to turn on its lights and to brake stronger to avoid a collision. Cars can take plenty of decisions today. We entrust our life to a machine that is much more sophisticated than any available robot. But today society is more afraid of robots than of cars. And there are less ethical questions about cars than about robots.

The way society managed to deal with automobiles, that have killed and will kill many more people than robots, is issuing drivers licenses. Before being allowed to drive the complex object that a car is, the driver is supposed to be taught and an evaluation is performed to be sure he will not become a risk to the other people on the road. To deal with the complexity of a robot, the user will have to have certain knowledge on the way the robot works, behaves and reacts. The education on robotics is necessary to insure that men are controlling the machine and not the contrary. Maybe a robotic license should become mandatory to use safely a sophisticated robot.

References

Albu-Schaeffer A, Eiberger O, Grebenstein S, Haddadin S, Ott C, Wimboeck T, Hirzinger G (2008) Soft robotics: from torque feedback controlled lightweight robots to intrinsically compliant systems. IEEE Robot Autom Mag:20–30

Asimov I (1950) I, robot. Gnome Press
Bemelmans R (2015) Effectiveness of Robot Paro in intramural psychogeriatric care: a multicenter quasi-experimental study. J Am Med Directors Assoc:946–950
Devillers L (2015) Inference of human beings' emotional states from speech in human-robot interactions. Int J Social Robot:451–463
Doan N (2015) High speed running of flat foot biped robot with inerter using SLIP model. In: Proceedings of advanced intelligent mechatronics (AIM), pp 110–115
Feng S (2015) Optimization based controller design and implementation for the Atlas robot in the DARPA Robotics Challenge Finals. In: Proceedings of humanoids, pp 1028–1035
Gouaillier D (2009) Mechatronic design of NAO humanoid. In: Proceedings of ICRA, pp 769–774
Ishiguro H (2015) The future life supported by interactive humanoid. IEEE Transducers
Kaneko K (2015) Improvement of HRP-2 towards disaster response tasks. In: Proceedings of humanoids, pp 132–139
Mukai T (2010) Development of a nursing-care assistant robot RIBA that can lift a human in its arms. In: Proceedings of IROS'10, pp 5996–6001
Pandey A (2015) Developmental social robotics: an applied perspective. Int J Social Robot:417–420
Riddle P (2015) http://projects.laas.fr/riddle/
Romeo P (2015) http://projetromeo.com/
Sandry E (2015) Re-evaluating the form and communication of social robots. Int J Social Robot:335–346
Tisseron S (2015) Le jour où mon robot m'aimera: vers l'empathie artificielle
Vidal D (2012) Vers un nouveau pacte anthropomorphique ! Les enjeux anthropologiques de la nouvelle robotique. In: Vidal D (ed) Robots étrangement humains, Gradhiva, pp 54–75

Chapter 17
Robots in Ageing Societies

Maria Isabel Aldinhas Ferreira and João Silva Sequeira

Abstract Addressing the topic of ageing societies, the present paper stresses the importance of preserving the autonomy, social participation and affective bonds of elders. Claiming that maintaining the social and affective ties that link someone to their home environment and to their close family and friends is fundamental for physical and mental health, and consequently for extended years of life with quality, the authors identify the potential benefits of assistive/domestic robots adverting to the potential ethical issues to be safeguarded.

Keywords Active ageing · ICT · Social robotics · Ethics · Human rights for older persons

17.1 Introduction

Although countries are not ageing at the same pace, demographic trends reveal a very significant world changing age distribution resulting from increased average longevity and the deep decline in fertility rates (Fig. 17.1).

According to the United Nations Report United Nations (2002) in 1950, there were 205 million persons aged 60 or over throughout the world. At that time, only 3 countries had more than 10 million people 60 or older: China (42 million), India (20 million), and the United States of America (20 million). Fifty years later, the number

M.I. Aldinhas Ferreira (✉)
Centre of Philosophy of the University of Lisbon,
Faculdade de Letras of the University of Lisbon, Lisbon, Portugal
e-mail: isabelferreira@letras.ulisboa.pt

J.S. Sequeira
Instituto Superior Técnico / Institute for Systems and Robotics,
Universidade de Lisboa, Lisbon, Portugal
e-mail: joao.silva.sequeira@tecnico.ulisboa.pt

M.I. Aldinhas Ferreira et al. (eds.), *A World with Robots*,
Intelligent Systems, Control and Automation: Science and Engineering 84,
DOI 10.1007/978-3-319-46667-5_17

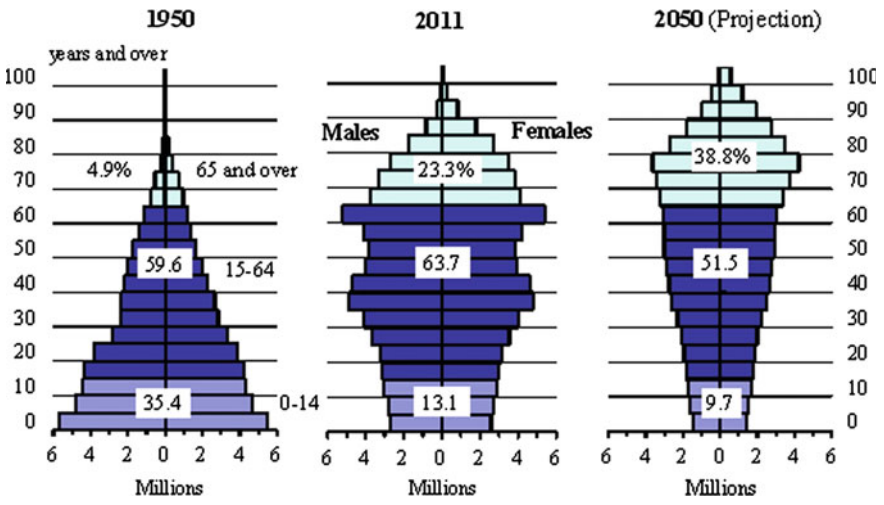


Fig. 17.1 Changes in the world population pyramid (from Statistics Bureau, MIC, Japan (2010))

of persons aged 60 or over increased about three times to 606 million. In 2000, the number of countries with more than 10 million people aged 60 or over increased to 12, including 5 with more than 20 million older people: China (129 million), India (77 million), the United States of America (46 million), Japan (30 million) and the Russian Federation (27 million).

Worldwide, the proportion of people age 60 and over is growing faster than any other age group being expected to reach nearly 2 billion in 2050.

Population ageing has profound consequences on a broad range of economic and social processes as most developed countries have "Pay-as-you-go" pension schemes, where publicly-provided pension benefits to older persons are paid out of taxes and social security contributions from current workers. The increased average longevity of senior citizens and the deep decline in fertility rates, consequently, ask for specific policies in order to preserve, on one hand, the quality of life of the elderly—at present mainly guaranteed by the public health care and social care systems—and on the other hand the stability and sustainability of the whole economic framework where the gap in the ratio pensionists/working force progressively increases.[1]

An ageing population raises challenges not only from the economic but also from the social and cultural points of view. In the last decades we have progressively assisted to the erosion of the intergenerational dependencies that traditionally guaranteed the care and affective support of the elderly by their younger close relatives. Migratory patterns within Europe, reflecting the movement (largely that of young

[1] According to the European Commission—The 2012 Ageing: Economic and Budgetary Projections for the 27 EU Member States (2010–2060), the balance of active (working) versus inactive people will shift from 5:1 to 2:1 in 2030.

people) from rural to urban areas, or the movement to more promising countries, have resulted in a reduction of traditional extended family structures.

This erosion of the longstanding family ties associated to the additional years of life that the most recent cohorts have gained (and stand to gain), as a result of declining mortality rates, is translated in an increasing number of aged people being left in residence houses, frequently for periods as long as twenty years, generally nearly cut off from any family and/or social ties, living a dull and psychologically painfully routine that is at most cheered up by inconsequent time-killer activities.[2]

On the other hand, those that choose to go on living in their houses and original communities are also often progressively left alone and unprotected.[3]

In this context, getting older signifies, for most people, the absence of socially assigned roles, a substantial decrease of reivindicative power, the progressive loosing of all social and affective ties, the cutting off with the real world and consequently a substantial loss of identity, Ferreira (2011).

This scenario raises many questions not only concerning the sustainability of a progressively heavier health and care systems but essentially about the actual individual and collective benefits of extended life expectancy—re the additional years of life that the most recent cohorts have gained years of good health or just years of disability and frailty? Howse (2006).

17.2 Changing Paradigms: The Active Ageing Policy

Aware of the ethical issues raised by the scenarios described above and acknowledging the importance for the individual's health of preserving the relationship individual/community and the net of social and affective links it involves, The United Nations Principles for older People, resolution 46/91 recommended that the elder go on living integrated in their communities preserving the net of social roles and emotional links that gave and go on giving significance to their existence.[4]

The concept of "Active Ageing" adopted by the World Health Organization (WHO)[5] in the 90's claims that individuals have the right to participate in society according to their motivations, will and capacities independently the phase of life they are in.

The WHO stresses that "active" refers to the continuing participation in social, economic, cultural, spiritual and civic affairs, not just the ability to be physically

[2]Epidemiological studies (The Medical Research Council Cognitive Function and Ageing Study in Medical Research Council (2010) have indicated the prevalecence of depression in institutionalised aged responsible for cognitive and physical decay.

[3]Among the European countries, Portugal is the fourth with the highest percentage of old people. According to the data released for Portugal-Census 2011, 20 % of the Portuguese population is over 60. About 1.200.000 old people live by themselves or in the company of another older and about 75000 live in residence houses, a number that tends to increase.

[4]World Population Ageing: 1950–2050. United Nations Report, 2002.

[5]Active Aging: a Policy Framework. April 2002. World Health Organisation.

active or to participate in the labour force. As it is pointed out, ageing should take place within the context of friends, work associates, neighbours and family members in a framework where the concepts of interdependence and intergenerational solidarity are fundamental tenets. Older people who retire from work, are ill or live with some disabilities should remain, according to their capacities, active contributors to their families, peers, communities and nations while these should provide them with the adequate protection, security and care when they need.

Over the past two decades, the "active ageing" trend has emerged in Europe as the foremost policy response to the challenges of population ageing. Within this framework a particular emphasis is placed in the role played by Information and Communication Tecnhologies (ICT).

17.3 Technological Development: An Enhancer of Satisfaction, Sustainability and Success?

The European Commission defined as a goal to help senior citizens to lead independent, healthy and productive lives for as long as possible, continuing to enjoy the comfort of their own homes. Its aim is to achieve a triple-win through ICT-based innovation: (1) more years of active and independent living for older people; (2) increased efficiency and sustainability of the health and social care systems; (3) promotion of a flourishing innovative industry enabling Europe's economic growth.

Over the last years, research projects in robotics for ageing well have been funded under the ICT strand of the seventh research framework programme (FP7) and under the Active and Ambient Assisted Living joint programme (AAL), with a total budget of 50M€. Under the new Research and Innovation Framework Programme Horizon 2020 (H2020) a batch of care robotics projects have been launched in the first quarter of 2015, with a total funding amount of 185M€.

The reinforcement of the Active and Ambient Assisted Living Programmes—jointly financed by the EU and 23 European countries[6]—has been a major element of the EU policy action on ICT for Ageing Well, being the focus, currently, on how to make robots cooperate and communicate with people in the context of intelligent environments. Its scope is broad comprehending smart houses with sensors for fire burst or gas or water leaks and/or detection of people's fall; appliances capable of monitoring the individual's health conditions, such as blood pressure, heart rate or diabetes level and being able to alert emergency systems whenever necessary; robotic systems capable of assisting in the performance of daily tasks such as taking a medicine, eating or exercising; or capable of carrying out difficult tasks such as moving a heavy object; rich interfaces capable of providing permanent cognitive stimuli or serving as communication systems to connect people with their carers, their doctors or family and friends.

[6]See Ambient Assisted Living Programme (2015) and Active and Assisted Living Programme (2015).

Fig. 17.2 The robot (butler) confirms the request and then switches off the lights

Within this framework, RoCKIn an EU-funded project[7] aims to foster scientific progress and innovation in cognitive systems and robotics through the design and implementation of competitions. In RoCKIn@Home the challenge focuses on assisting the elderly or impaired. Here the robot acts as a buttler, moving about at the command of the user, switching on and off lights or electronic systems with a voice command, or using its arm to fetch things, answering the front door and keeping track of the current position of a person from a different place in the house (Fig. 17.2).[8]

17.4 Some Ethical Concerns

Roboethics is the ethics inspiring the design, development and employment of Intelligent Machines Verugio (2006). Roboethics is about human ethics and consequently must respect fundamental human values.

The main objections that have been raised about the use of robotic technology with vulnerable older people are (i) loss of social contact (ii) deception and, in some cases, even (iii) infantilisation.[9]

Concerns about the possible loss, or reduction, of the amount of social contact for older people, following the introduction of robot pets, were also raised by Sparrow and Sparrow (2006). Their argument is that older people are likely to suffer a reduction in the amount of human contact as more robots are introduced in their

[7] http://rockinrobotchallenge.eu.

[8] Catering for Granny's house video reflects the work of ISR/IST researchers within this project https://www.youtube.com/watch?v=a_F0xMyyJ0s.

[9] Cf Sharkey and Wood (2014).

daily environment. They suggest that even the introduction of a floor cleaning robot could result in the loss of the potential social interaction with the human cleaner it replaces.

Concerns are often also expressed about the importance of maintaining the dignity of people as they age. Though there is a certain lack of clarity and agreement about what the term 'dignity' actually means, this recommendation, present in reports such as the 2012 National Pensioners Dignity Code or The Chart of Human Rights for Older Persons Council of Europe (2014), clearly identifies the need to respect older individuals, their life options and constructed environments as well as treating them with all the respect and formality due to any adult.

Though one can easily anticipate the beneficial contribution played by the social/assistive robotics technology extending the autonomy level of elder and thus allowing them to keep living at home, its introduction in the domestic environment on a daily basis is a very complex issue.

In our opinion, three inalienable human rights have to be safeguarded:

1. The right to freedom, in all instances, to make their own choices, e.g.,
   - The choice to live or not in a smart environment where personal status and routines can be monitored
   - The choice to share this environment with a robot- it's not the same interacting with a robot occasionally than sharing the same environment with a physical entity 24 hours a day, listening to its frequent or continuous prompts.
2. The right to privacy and intimacy at physical at psychological levels
3. The right to real interaction with other humans and the inalienable right to love and be loved
   - Sharing their environment with close relatives and friends
   - Always being taken care by physically present human subjects
   - Not being deceived and lead to conclude their artificial addressee is an animate entity endowed with the capacity to care and eventually to love

17.5 Conclusions

Robotic technology can benefit the quality of life of older or impaired people and by contributing to their extended autonomy and health status enhance a more balanced economic framework. However the impact of its deployment in domestic and in institutional environments must be carefully anticipated.

As it happens with all technology, robotic technology has to follow a user-centred approach:

(i) Identifying and distinguishing the profiles of end-users: their cultural contexts, typical environment, their needs, preferences and expectations.
(ii) Producing technology that is safe, reliable and not intrusive easily adapting to the physical environment and to the conditions of the user.

(iii) Never leading into deception.
(iv) Never replacing the direct interaction with other humans and the sharing of feelings and emotions.

Perhaps the most serious threat lies in the possibility of replacing real social and affective bonds by close relatives, friends and carers with the interaction provided by the machine.

In fact if we consider the initial scenario from the film "Robot and Frank", the American science fiction comedy-drama film, Jake Schreier (2012), where the son gives his father a robot so that it provides him with company and help, we can easily see how technology can progressively come to assure the "easiness" of consciousness of those that primarily should not only be responsible but fundamentally love, care and preserve their close family ties.

But maybe robotic technology also gives a chance to rethink the present societal organization allowing to build an even more human society.

References

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

Ambient Assisted Living Programme (2015). http://www.aal-europe.eu/

Council of Europe (2014) The Chart of Human Rights for Older Persons. https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=09000016805c649

Ferreira M (2011) On meaning: individuation and identity. Cambridge Publishers, UK

Howse K (2006) Increasing life expectancy and the compression of morbity: a critical review of the debate. In: Kreager P (ed) Oxford Institute of Ageing Working Papers, University of Oxford. http://www.ageing.ox.ac.uk

Medical Research Council (2010) The Medical Research Council Cognitive Function and Ageing Study. http://www.cfas.ac.uk

Schreier J (2012) Robot & Frank. Science fiction comedy-drama film, directed by Jake Schreier. Samuel Goldwyn Films

Sharkey A, Wood N (2014) The Paro seal robot: demeaning or enabling? http://doc.gold.ac.uk/aisb50/AISB50-S17/AISB50-S17-Sharkey-Paper.pdf

Sparrow R, Sparrow L (2006) In the hands of machines? The future of aged care. Minds Mach 16:141–161. http://npcuk.org/710, http://npcuk.org/wp-content/uploads/2011/11/NPC-Dignity-Code.doc

Statistics Bureau, MIC, Japan (2010) Statistical Handbook of Japan 2010. http://basementgeographer.com/rural-aging-in-japan-and-population-implosion/, http://www.stat.go.jp/english/data/handbook/index.htm

United Nations (2002) World Population Ageing: 1950–2050. http://www.un.org/esa/population/publications/worldageing19502050/, United Nations Report

Verugio G (2006) Roboethics: a Bottom-up interdisciplinary discourse in the field of applied ethics in robotics. Int Rev Inf Ethics 6. http://imaginary.org/

Part II
Associated Events

"At the bottom Robotics is about us. It is the discipline of emulating our lives, of wondering how we work."

Rod Grupen (2008)

Chapter 18
Nós e Os Robots/Os Robots e Nós: Insights from an Exhibition

Rodrigo Ventura and Maria Isabel Aldinhas Ferreira

Abstract Being an integrant part of ICRE 2015, the exhibition "Nós e os Robots/Os Robots e Nós", "We and the Robots/The Robots and Us" aimed to bring robotic technology closer to the average citizen in an educational effort that, we believe, should precede the massive deployment of all Information and Communication Technologies and that becomes particularly needed at the verge of a widespread use of robotic technology. This paper gives a brief account of the content and organization of that exhibition and of how the public reacted to it.

Keywords Robotic technology · Social robots · HRI · User's awareness and education · Ethics · Harmonious integration

18.1 Introduction

The increasing development of robotic technology in the last decades is leading to its fast implementation in distinct domains of human life. Factory automation, transportation, military purposes, medical appliances and service applications for edutainment or personal assistance are just some of the fields where this technology has been deployed.

Robotic technology was initially restricted to the secluded industrial environments (e.g., Fig. 18.1a) or to the scenarios of research testbeds. In both cases none or very

R. Ventura (✉)
Institute for Systems and Robotics, Instituto Superior Técnico,
Universidade de Lisboa, Lisbon, Portugal
e-mail: rodrigo.ventura@isr.tecnico.ulisboa.pt

M.I. Aldinhas Ferreira
Centre of Philosophy of the University of Lisbon,
Faculdade de Letras, University of Lisbon, Lisbon, Portugal
e-mail: Isabel.ferreira@letras.ul.pt

M.I. Aldinhas Ferreira et al. (eds.), *A World with Robots*,
Intelligent Systems, Control and Automation: Science and Engineering 84,
DOI 10.1007/978-3-319-46667-5_18

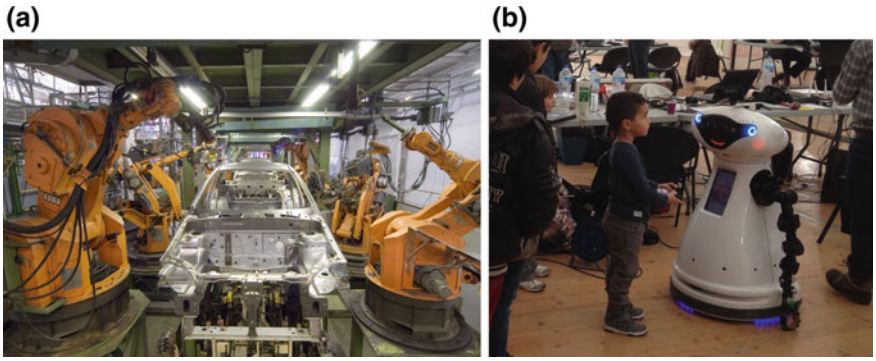


Fig. 18.1 **a** Robots in an industrial environment, KUKA (2016); **b** Robots interacting with humans, RoCKIn (2016)

limited interaction with common people or even visibility was expected. However, in the last decades, robots have progressively become involved in increasingly more complex and less structured tasks and activities, having started to interact with real people in public and/or private social spaces (e.g., Fig. 18.1b).

Robots are designed by people to be used by or with people, in human benefit, this means, effectively contributing to the well-being and ultimately happiness of human beings. This is, in fact, what it is generally expected of technological development, as a whole, though all along the road, this is not always a fact.

However, successful deployment does not just depend on the efficiency and effectiveness of its performance but also on its harmonious integration in human life. This integration demands the creation, before its massive implementation, of space and time for people to come close, get acquainted and have the chance to interact with the different forms robotic technology can assume. Giving them the chance to come to know more about the technology and its limitations, to express their doubts, to learn how to interact with it and last but not least to develop an awareness relatively to the ethical issues that can already be anticipated. This is an educational process that should be set as a priority in all forms of technological development, but that is particularly relevant to the ones with huge impact on the way people live and establish relationships, as it happens with ICT technologies.

The exhibition "Nós e os Robots/Os Robots e Nós" that took place in the main exhibition room of the Pavillion of Science, in Lisbon, was a fundamental part of ICRE 2015. The goal of the exhibition was to create a space and time for people to come close to real technology, interacting with it and starting to make an idea on how it will be when robots come to share our daily environment in a constant way.

The exhibition's title aims to highlight the circular nature inherent to the semantics of the concept of (interaction) itself. The virtual circle drawn by every instance of HRI, binding the end-user and the robot, is a process in which both parts are assigned an identical agent role. The Portuguese and French morphology and syntax reflect this symmetrical relationship: "Nous et les Robots/Les Robots et Nous", "Nós e os

Robots/Os Robots e Nós" translates a relationship where each of the participants is, at a time, respectively, the subject and the recipient in the closed circle that every interaction always involves.

18.2 Exhibition and Exhibitors

As mentioned before, the exhibition's main goal was to bring close to the public a broad range of real autonomous robots performing live demos. These robots represented not only research and development activities of the academia, but also commercial applications promoted by companies with robots as their core products.

The Institute for Systems and Robotics (ISR/IST), a university-based research center focused on advanced multidisciplinary research activities, in the areas of Robotic Systems and Information Processing, participated in the exhibition with the works of three of its research groups.

The Dynamical Systems and Ocean Robotics Laboratory (DSOR), showing videos of their marine autonomous vehicles (Fig. 18.2), since it was not possible having them performing there, the Computer and Robot Vision Laboratory (VisLab), demoing Vizzy, Fig. 18.3a, a wheeled humanoid robot for assistive robotics performing simple tasks, such as people following and object grasping, and the Intelligent Robots and Systems Group (IRSg), making demos of the wheeled robot MBot (Fig. 18.3b) performing tasks in a domestic setting, integrating speech interaction, navigation, and object manipulation.

Another university-based research group focused on autonomous agents- the Intelligent Agents and Synthetic Characters Group (GAIPS) from INESC-ID, presented several interactive demos involving embodied agents that interact with the public by involving them in games. One of the demos had a large interactive table through which the public could play a game with a robot. The interaction was affective, in the sense that the human player's emotions were explicitly modeled by the system.

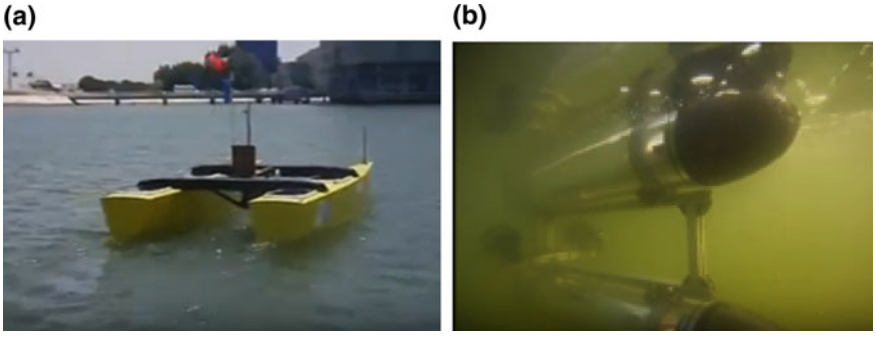


Fig. 18.2 **a** DELFIMx (2016); **b** Medusa (2016)

Fig. 18.3 **a** Vizzy (2016); **b** MBot (2016); **c** Frog (2016)

IdMind, a IST spin-off devoted to the development and construction of mobile robots, demoed several robots, including FROG (Fig. 18.3c), an outdoor guide for public spaces, such as a zoo.

It should also be mentioned that the Pavilion of Knowledge, where the exhibition took place, is currently using a mbot as a welcoming guide to the museum. This robot, called Viva, was also circulating among the exhibition area, actively interacting with the public.

TURFLYNX was another IST spin-off present. This is devoted to the development and construction of autonomous lawn mowers for large areas. Since their robots were too big to fit inside the exhibition building, they were present with videos and informational leaflets.

Being a company that exploits the intersection between robotics and art, Artica, a company specialized in multimedia interactive installations, demoed Gyro, a small affordable robot built on-top the Arduino platform.

The Portuguese Air Force Academy (AFA), also present, has been very active in the past years on the development and construction of drones for defense applications, namely for surveillance missions. Having developed the capacity for building these drones from scratch up to the systems integration and validation, they showed some of their drones, together with videos featuring the missions they have been involved recently.

Together, these exhibitors covered not just several dimensions of the presence of robots in our society- from completely non-interactive applications to social/service robots capable of exhibiting affective behaviors- but they brought together research prototypes and commercial applications (Figs. 18.4 and 18.5).

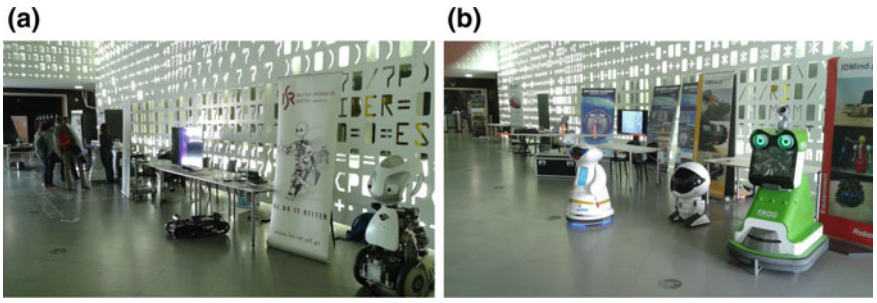


Fig. 18.4 Photos of the exhibitors: **a** from *left* to *right*, GAIPS and ISR, **b** from *left* to *right*, TURFLYNX, Artica, AFA, and IdMind. The Viva robot is also visible in **b** (*leftmost* white mobile robot)

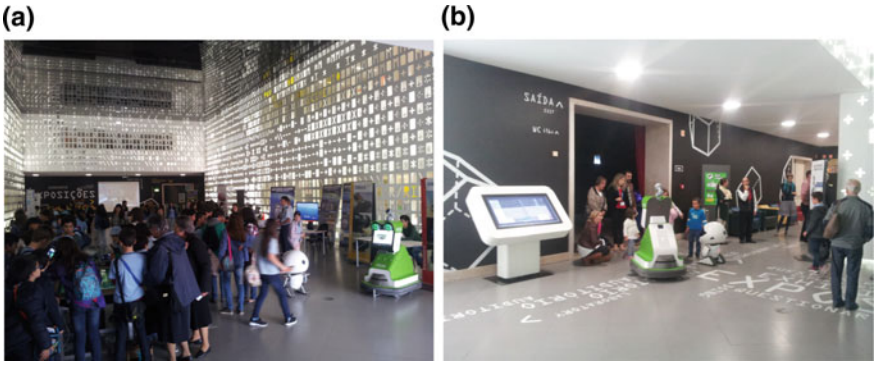


Fig. 18.5 Interacting with the robots

18.3 The Reaction of the Public

The exhibition was visited by groups of children from schools on working days and by whole families during the weekend. The children were extremely enthusiastic with the mobile platforms. This was pushed to the extreme once they were allowed to remote control the robots: they formed queues waiting in line for the chance of remote controlling a robot, even for a few seconds each.

It should also be noted that, even thought the behaviors exhibited by these mobile robots were extremely simple, children very often ascribed to them much more complex constructs, such as intentions and emotions. This contributed to their engagement as well as to the active participation of adults who were also very curious about the robotic technology. In most cases, the exhibition provided a great opportunity not only for a first interaction with robots, but also for lively discussions between the public and the robot manufacturers about their technological specificities but also about their roles in daily life.

18.4 Conclusion

The interest and curiosity shown by grown ups and children relatively to the robots on display and their excitement when interacting with them, during the exhibit, demonstrates the educational role this type of events can play. We firmly believe that the periodic realization of events of this kind provides a fundamental educational tool in what relates the end-users' future practices, guiding them on the best ways on how to benefit from technological development in a framework where safety and ethical standards are respected and user-experience is gratifying.

References

DELFIMx (2016) https://www.youtube.com/watch?v=CHj13WdB3B4
Frog (2016) http://querosaber.sapo.pt/tecnologia/frog-o-primeiro-robo-guia-chega-a-portugal
KUKA (2016) https://commons.wikimedia.org/wiki/File:KUKA_Industrial_Robots_IR.jpg
MBot (2016) ISR/IST—University of Lisbon. https://www.youtube.com/watch?v=baXkyNAdXm4
Medusa (2016) https://www.youtube.com/watch?v=8GJjzUCrcYU
RoCKIn (2016) Rockin Project—irs/ist University of Lisbon
Vizzy (2016) Vislab—ISR/IST-University of Lisbon

Chapter 19
The Robot Steps In: From Normative to Prospective Ethics

José Manuel Martins

Abstract The present paper reports the role played by the cinema cycle associated to the International Conference on Robot Ethics (ICRE 2015), that was open to the general public. Reflecting on the grounding motivations leading to this international conference, the paper analyses the role played by fiction and film industry in the definition and anticipation of a future world where humans and intelligent machines will coexist.

Keywords Fiction · Robots · Ethics · Societal models

19.1 Introduction

Golem, or 'Frankenstein', are no longer those portentous automata filling in a fictional place within the proto-industrial collective imaginary haunted by regressive magical features. Neither do they represent as many historical forebodings and cautionary tales at the ominous crossroads of Modernity. Over the course of time, they became the actual robots we come across everywhere in the real world of industrial environments, entertainment facilities and advanced technological expertise—and they are on the brink of invading our daily and domestic life, too.

In this context, what can be the role of fiction in elucidating this ideal of an ultra-technological improvement of the Robot—at first, an ideal of ours, then, eventually, of its own in the dazzling era of the increasingly real possibility that it occurs?

The International Conference on Robot Ethics—ICRE 2015—significantly invoked on its website the fictional legacy of that founding father of robotic literature

J.M. Martins (✉)
Department of Philosophy, University of Évora Center of Philosophy
of the University of Lisbon, Lisbon, Portugal
e-mail: jmbmarte@gmail.com

M.I. Aldinhas Ferreira et al. (eds.), *A World with Robots*,
Intelligent Systems, Control and Automation: Science and Engineering 84,
DOI 10.1007/978-3-319-46667-5_19

that was Isaac Asimov by quoting his famous 'Three Laws of Robotics', whose explicit formulation dates back to the 1942 novel Runaround.

And ICRE even inscribed that futurizing and thought-provoking product of a science-fictional speculation at the very frontispiece of an academic event, focused on reflecting on the ethical contexts of human/robots interactions, thus unifying the internal functionalities and the operating contexts, both physical and ethical, under a single design and ethical principle.

19.2 From Normative to Prospective Ethics: The Role of Fiction

In both industrial and educational, entertaining, military or healthcare robotics, the ethical bond defining the realm and the modalities of a 'robotic ethics' remains the same that rules reciprocal responsability between humans regarding the presence and the interference among them of a peculiar operating system enjoying increasing mobility and performativity, that must (or ought to) serve without causing harm.

There is no question of an ethic(s) of the human towards the robot(ic), despite the appearance that an inchoative robot-towards-humans ethics (vg, to avoid hurting us) would seemingly have been already encrypted into its operating systems as an automatism. In fact, though, such an ethics is but the projective transposition of a strictly inter-human ethics (one that forbids, let us say, to build dangerous toys) into the robot's automatic system's design. Whenever blind dangerousness is restricted through specific mitigating circuits, these are conspicuously not an ethical quality—a 'virtue'—of 'the robot itself' towards the human, but the simple mechanical and electronic inscription of an ethical intention of the human manufacturer towards his fellow human customer, through an interposed robot. The ethical perimeter is inscribed into the robot not as an ethical dimension (the robot does not 'become cautious'), but as a network of automatic performance circuits (the programmer ensures that the mechanical movements of the apparatus be compatible with human co-working contexts). In the case of the military drone, governed from afar or pre-programmed, the ethical question becomes once again, and more poignantly than ever, a human question—and certainly not a robotic one.

Robot ethics is neither the ethics of robots (among themselves, towards themselves and/or towards us) nor the (human) ethics towards robots, because precisely there is no such thing as 'the robot itself.' Robots are object-like entities, not subjects, at least for the time being, and as far as we can see. And it is the time being, and seeing, that cinema is all about.

What is the impulse, then, deliberate or inscrutable, that drove ICRE 2015 to introduce the fictional ethics of Asimov's robotics within the scientific agenda of a congress whose concerns thus leap over the state-of-the-art in this domain and hint at a 'robot(ic) ethics' as one that would concern robots as such?

We may indeed enumerate, in the title expression and on the front page arrangement chosen by ICRE, three successive semantic intentions and ranges:

1. The exploration of an applied ethical field, the human/robot interface, as a particular instance of the human/machine or human/technology interfaces, stemming from the increasing (and combined) cyber-mechanical autonomy attained in robotics, which qualitatively surpass, on one side, the immobile mechanic automation of the classical industrial apparatuses and assembly lines (in addition to that, robots do move, and they 'think' better); on the other, the cybernetic processing of the typically virtual and nonlocal artificial intelligence (again, robots not only move, they also sense, and are designed to interact 'culturally').
2. The suggestion that the introduction of robots in the human world gave rise to an ongoing process, sliding from a purely human-centered ethics into a sort of robot-wise ethics (or into a re-categorization of robot behaviour as human-like subjecthood), whereby former basic ethical assumptions started being revised, far beyond a mere extension of the field of applied ethics. For instance, through the transference of human ethical properties to the (broadly generalized) care practices and acts carried out by robots, which would thus assume humanized functional features through programming; or, to say it otherwise, which would make an outwardly humanly operated input into their own ethical being. Robots would henceforth be acknowledged by their direct human beneficiary within a care practice context as actually "attentive, responsible, responsive, competent" in themselves and by themselves, and not just endowed with (transposed) "attentiveness, responsibility, etc.". In the grammatical ontology of this ethics, robots appear as the subjects of their own verbal acts ("look how attentive and caring this little robot actually is!"), and not just as the carriers of abstract marks metonymically attributed to it: "see how smartly the attentiveness circuits of this service robot have been devised!". Still, in both cases, far from self-asserted, the ethical sphere remains dependent on an onlooker who is able to (mis)recognize it as such. However, it is precisely this tiny frontier that we are committed to explore alongside our movies.
   From yet another angle, we can observe that a new ambiance is brought to bear on the human ethos by the shared abidance by the common security rules governing both humans and robots in v.g. an industrial working context: they are conjointly mobilized and subsumed as interfacing entities under a new ethics of posthuman cohabitation imparting modified behaviour patterns.
3. Finally, we must register the willing quotation of the Asimovian 'Three Laws of Robotics', which apply properly to fictional beings infinitely more advanced then the current robots: to beings on their way to becoming full-fledged subjects—either the subjects of their own emancipated robotic identity, or subjects of the identity after which they ambiguously model themselves: human identity.

That is precisely the point: on their way to, and in exactly the state of ambiguity where the nomothetic status of those "laws" may mean three different things:

(α) physical laws: a structural and constitutive property—unavoidable and indecomposable—of the very logical-mathematical bedrock of the Asimovian fictional 'positronic brain' operating as plain mechanism;
(β) logical laws: the insertion of a circuit board containing additional or optional algorithms determining automatic safeguard operations (coincident with the sort of 'users safety' programmers ethics currently recommended with regard to robotic devices);
(γ) ethical laws: the proposed inscription of moral obligation protocols, at once binding and free—similar to the engraving of the Ten Commandments on the holy stones or of the Kantian maxim into the rational consciousness of the moral beings—, onto the assembled electronic web, presuming the agency of an emergent self-consciousness able to assume them.

What sort of unsettling ambivalence, if any, is thus surfacing in this hybridization between the fictional and the 'actually real'?

We would risk to say an ambivalence parallel to that of those three competing understandings of the status of the Three Laws, that imply speculatively an improvement in robotics (hence, in the robot) escalating from assembling a deterministic organism to birthing an autonomous selfconscious being, 'below and beyond the human'. It all comes to acknowledging that the fictional creation of the robot, both in literature and in cinema, while closely anticipating its actual technological production in the course of history, voices an imaginary desire for perfectiveness which is not less integral to the actual fabrication of robots according to a technological desire for perfectiveness that results in a fabrication of the very perfectibility of the robot. Meaning, that differently from manufacturing cars or coffee machines, addressing the sheer possibilities as such of a medium term exponential improvement of the product, is as much a primary goal of a robotics manufacturing program as the actual building of any 'current generation' of robots. The steps leading to the next level (or 'generation') along the perfection line of the robotic product are by definition pre-contained in the actual manufacturing taking place at any stage and serving at the same time as a test of the overall advancements in the field (each stage being put intrinsically into a calculative relation with the next ones by the recurring, exulting adverb 'already', referring to the degree of functionality 'already' attained). The inward nature of robots is, indeed, perfectibility—and evolution: the inversion of the 'already' into 'not yet' denotes that a (meta)manufacturing of its own future is part of every actual course of manufacturing a given generation of robots. These, in turn, have no equivalent amongst the production of artificial beings, for they embody this unique bio-mechanical-electronic entity that will ultimately converge, and in fact always-already has from the inception, with the ontological identity of the manufacturer herself.

ICRE-2015 cinema cycle, comprehending the films *Westworld—Blade Runner—I, Robot—A.I. Artificial Intelligence*, offered the possibility of going across the technological phenomenon of the generation of robots towards the fictional-real horizon of the generations of replicants which in turn come to meet the cyborgs- we on our side

are turning into- in the vertigo of a strangely convergent co-evolution determined by the unstoppable logic of present-day technology.

Every trait of physical and/or mental, ethical and 'emotional' anthropomorphization currently lent to the image of the robots' quidditative being points ('already') in that direction and indicates that the obscure Desire is 'fictionality' inscribed, in the present case, within the very nature of the reality at stake. A Desire of 'reassembling' the human mistery? Of a divine duplication of the human creature's lonely face? Of potentiating and controlling a super-human Power, so delegated as to be exerted with impunity?

Techno-evolutionary reality incubates deep inside itself a drive towards exponential mutability. A rule of an unheard-of progression, which the so called 'science-fiction' first brought to light as a forewarning, then went on disclosing as a literary realm of its own, and now explores 'in real time' alongside the actual and the expected progress of techno-science. Now, this same drive also presides explicitly over the logic of the technological and scientific research in those realms: deciphering the code of life, cybernetic modeling consciousness, developing artificial intelligence and combining it with biotechnology until obtaining the optimized (and ultimately indistinguishable) technological replica of the human model thus decoded and completely restored.

The current robot is a prefiguration of its ingress into the human (a prefiguration which is functional and effective as well as wishful and fantasized; a vehicle for material efficacy as well as for projective dreams). The robot's ingression into the human condition overlaps the return of humanity unto itself: and it is that pathos of the recognition of oneself in the Other that governs emotionally the sentimental shock that bends us into taking the side of the androids in all those four films—that is to say, the side of ourselves, within our deepest longing for what we are. An aspiration to fulfill that ideal of humanity so transcendent that, at the same time it keeps escaping us, also became the redemptive ideal for a being we created in order to be more than itself. A being which we, in the ongoing state of our evolution in designing its, call a 'robot'.

The Slavic etymology of the word—'robot'—provides the semantic and political bridge between an all too human sense of the age-long slavish worker and the Jüngerian figure of the 'total mobilization' of the modern Worker, but also between these and their mechanized counterpart. The Worker is the one whose essence strictly consists of, and reduces itself to, the mechanical act of working, and nothing else; the one whose being, thus deprived or emptied of itself and of any ethical dimension, while at the same time endowed with a consciousness (a being-for-itself), enacts a fundamental conflict already seminally theorized by Hegel and Marx and resurfacing in the overdetermined character of these four cinematic fables (as of so many others), oscillating between a humanistic ethical stand and an acrid political hue.

The role of fiction is not, then, to represent the (gaseous) imaginary region counterpointing the (solid) region of reality, according to a clear divide: the fictional/imaginary component of the 'Three Laws' (and of the four films) at this Conference does not play the usual role of presenting an 'imaginary variation' around the metaphysical 'possibles' of these or those other 'stable essence entities'—of

the type' a robot is. Instead, to represent the (fluid) process character of reality 'self-fictionalizing' itself: its character of an ongoing liquid mutability feeding the unstable 'in-between-ness' of our present day 'reality'.

Fiction inhabits and voices that very ontological self-displacement of the real. The imaginary is now the imaginary of the real, its generative segment. Fiction movies correspond to this moment of self-mediation of reality, they answer that robotic co-habitation already obscurely nurturing the prospective imagination of contemporary designers of robots/physical-cultural environments systems.

The four movies of ICRE's cinema cycle are not a delirium of the creative fantasy in the Seventh Art: they diagnose, rather (literally: they see through, they radiograph), a trend in contemporary techno-scientific evolution that not only anthropomorphizes robots and brings them closer to the human; but, in a reciprocal move, cybernetises humans and brings them closer to robots. The 'post-human condition' and the advent of the cyborg take place converging from both sides.

The history of the evolution of the Laws of Robotics—a fictional story of fictional laws—is, nonetheless, unwittingly enlightening about the whole ambiguity contained in the figure of the robot. Paradoxically, this later is all the more autonomous as it is more human, and thus harbouring in it a basic irremediable conflict that tends to escalate. Asimov formulated them as an antidote against the 'Frankenstein complex' (the monster out of control, or the spell turned against the sorcerer). And yet, the entire sci-fi realm went on since then, in every conceivable variation, to assert itself without a wrinkle as a cautionary tale about the universal takeover by the cybernetic world web, within a synergy between the 'mental' and virtual computer and the correspondent armed wing, the 'physical' robot.

But because the model that founded 'robotics' itself was the model of the political economy that reifies human subjects into labor tools or conversely, when object-like tools 'that look like a robot' become subjects, the conflict is no longer just ethical or legal, it is political. However, the political fable of robots does not stage allegorically our own, it does not commemorate Spartacus; neither does it represent a future robotized society, namely, maniacally structured upon work: it represents, rather, the matrix common to both of them—the economical-political and ethical matrix of what we are, that is to say: of who (?) we are.

Index

A
Accountability, 214
Ageing, 207
Airworthiness, 187
Automation, 160
Autonomous, 136
Aviation safety, 185

B
Bundled norm, 8

C
Caregiver, 78
Causal relations, 37
Causative agents, 124
Censorship, 68
Chisholm's Paradox, 51
Classification relations, 37
Cloud manufacturing, 160
Collaboration, 145
Confidentiality, 164
Contractual, 119
Cybercrime, 202
Cyber-physical, 160

D
Deontic operators, 6

E
Ethical concerns, 94
Ethical governor, 79
Evaluation relations, 38

F
Follower, 146
Free software, 63

G
Globalization, 201
Governance, 69

H
Humanitarian law, 174
Humanoid, 207

I
ICT, 220
Immunity, 128
Industrie 4.0, 160
Insecurity, 70
Integrity, 186
Intrusive, 210

K
Killer, 172

M.I. Aldinhas Ferreira et al. (eds.), *A World with Robots*,
Intelligent Systems, Control and Automation: Science and Engineering 84,
DOI 10.1007/978-3-319-46667-5

L
Leader, 146
Lethal, 171
Liability regime, 124

M
Machine ethics, 4
Machine morality, 4
Mediator, 78
Moral agency, 147
Moral norms, 5

N
Normative decisions, 36
Normative system, 36

O
Open source, 64

P
Parkinson, 78
Perceived autonomy, 21
Perceptual invariants, 141
Privacy, 71
Proactive, 210
Prohibition, 127
Publishing, 110

R
Redundancies, 186
Reliability, 186
Risk analysis, 191
Role, 146
Role switching, 148

S
Safety, 185
Search and rescue, 94
Social norms, 5
Spying, 71
Surveillance, 71

T
Terrorism, 201

V
Value sensitive design, 94
Value tensions, 94