SPRINGER BRIEFS IN ETHICS

Christoph Bartneck
Christoph Lütge
Alan Wagner
Sean Welsh

An Introduction to Ethics in Robotics and AI

Springer

SpringerBriefs in Ethics

Springer Briefs in Ethics envisions a series of short publications in areas such as business ethics, bioethics, science and engineering ethics, food and agricultural ethics, environmental ethics, human rights and the like. The intention is to present concise summaries of cutting-edge research and practical applications across a wide spectrum.

Springer Briefs in Ethics are seen as complementing monographs and journal articles with compact volumes of 50 to 125 pages, covering a wide range of content from professional to academic. Typical topics might include:

- Timely reports on state-of-the art analytical techniques
- A bridge between new research results, as published in journal articles, and a contextual literature review
- A snapshot of a hot or emerging topic
- In-depth case studies or clinical examples
- Presentations of core concepts that students must understand in order to make independent contributions

More information about this series at http://www.springer.com/series/10184

Christoph Bartneck · Christoph Lütge ·
Alan Wagner · Sean Welsh

An Introduction to Ethics in Robotics and AI

Springer

Christoph Bartneck
HIT Lab NZ
University of Canterbury
Christchurch, New Zealand

Christoph Lütge
Institute for Ethics in Artificial Intelligence
Technical University of Munich
München, Germany

Alan Wagner
College of Engineering
Pennsylvania State University
University Park, PA, USA

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

ISSN 2211-8101 ISSN 2211-811X (electronic)
SpringerBriefs in Ethics
ISBN 978-3-030-51109-8 ISBN 978-3-030-51110-4 (eBook)
https://doi.org/10.1007/978-3-030-51110-4

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

Fig. 1 The logo of the EPIC project

This book was made possible through the European Project “Europe’s ICT Innovation Partnership With Australia, Singapore & New Zealand (EPIC)” under the European Commission grant agreement Nr 687794. The project partners in this consortium are:

- eutema GmbH
- Intersect Australia Limited (INTERSECT)
- Royal Melbourne Institute Of Technology (RMIT)
- Callaghan Innovation Research Limited (CAL)
- University Of Canterbury (UOC)
- National University Of Singapore (NUS)
- Institute For Infocomm Research (i2r)

From February 2–6, 2019 we gathered at the National University of Singapore. Under the guidance of Laia Ros from Book Sprints we wrote this book in an atmosphere of mutual respect and with great enthusiasm for our shared passion: artificial intelligence and ethics. We have backgrounds in different disciplines and the synthesis of our knowledge enabled us to cover the wide spectrum of topics relevant to AI and ethics.

This book was written using the BookSprint method (http://www.booksprints.net).

Contents

List of Figures

Chapter 1
About the Book

This book provides an introduction into the ethics of robots and artificial intelligence. The book was written with university students, policy makers, and professionals in mind but should be accessible for most adults. The book is meant to provide balanced and, at times, conflicting viewpoints as to the benefits and deficits of AI through the lens of ethics. As discussed in the chapters that follow, ethical questions are often not cut and dry. Nations, communities, and individuals may have unique and important perspectives on these topics that should be heard and considered. While the voices that compose this book are our own, we have attempted to represent the views of the broader AI, robotics, and ethics communities.

1.1 Authors

Christoph Bartneck is an associate professor and director of postgraduate studies at the HIT Lab NZ of the University of Canterbury. He has a background in Industrial Design and Human-Computer Interaction, and his projects and studies have been published in leading journals, newspapers, and conferences. His interests lie in the fields of Human-Computer Interaction, Science and Technology Studies, and Visual Design. More specifically, he focuses on the effect of anthropomorphism on human-robot interaction. As a secondary research interest he works on bibliometric analyses, agent based social simulations, and the critical review on scientific processes and policies. In the field of Design Christoph investigates the history of product design, tessellations and photography. The press regularly reports on his work, including the New Scientist, Scientific American, Popular Science, Wired, New York Times, The Times, BBC, Huffington Post, Washington Post, The Guardian, and The Economist.

C. Bartneck et al., *An Introduction to Ethics in Robotics and AI*,
SpringerBriefs in Ethics, https://doi.org/10.1007/978-3-030-51110-4_1

Christoph Lütge holds the Peter Löscher Chair of Business Ethics at Technical University of Munich (TUM). He has a background in business informatics and philosophy and has held visiting positions in Harvard in Taipei, Kyoto and Venice. He was awarded a Heisenberg Fellowship in 2007. In 2019, Lütge was appointed director of the new TUM Institute for Ethics in Artificial Intelligence. Among his major publications are: "The Ethics of Competition" (Elgar 2019), "Order Ethics or Moral Surplus: What Holds a Society Together?" (Lexington 2015), and the "Handbook of the Philosophical Foundations of Business Ethics" (Springer 2013). He has commented on political and economic affairs on Times Higher Education, Bloomberg, Financial Times, Frankfurter Allgemeine Zeitung, La Repubblica and numerous other media. Moreover, he has been a member of the Ethics Commission on Automated and Connected Driving of the German Federal Ministry of Transport and Digital Infrastructure, as well as of the European AI Ethics initiative AI4People. He has also done consulting work for the Singapore Economic Development Board and the Canadian Transport Commission.

Alan R. Wagner is an assistant professor of aerospace engineering at the Pennsylvania State University and a research associate with the universities ethics institute. His research interest include the development of algorithms that allow a robot to create categories of models, or stereotypes, of its interactive partners, creating robots with the capacity to recognize situations that justify the use of deception and to act deceptively, and methods for representing and reasoning about trust. Application areas for these interests range from military to healthcare. His research has won several awards including being selected for by the Air Force Young Investigator Program. His research on deception has gained significant notoriety in the media resulting in articles in the Wall Street Journal, New Scientist Magazine, the journal of Science, and described as the 13th most important invention of 2010 by Time Magazine. His research has also won awards within the human-robot interaction community, such as the best paper award at RO-MAN 2007.

Sean Welsh holds a PhD in philosophy from the University of Canterbury and is co-lead of the Law, Ethics and Society working group of the AI Forum of New Zealand. Prior to embarking on his doctoral research in AI and robot ethics he worked as a software engineer for various telecommunications firms. His articles have appeared in *The Conversation*, the *Sydney Morning Herald*, the *World Economic Forum*, *Euronews*, *Quillette* and *Jane's Intelligence Review*. He is the author of *Ethics and Security Automata*, a research monograph on machine ethics.

1.2 Structure of the Book

This book begins with introductions to both artificial intelligence (AI) and ethics. These sections are meant to provide the reader with the background knowledge necessary for understanding the ethical dilemmas that arise in AI. Opportunities for further reading are included for those interested in learning more about these topics. The sections that follow focus on how businesses manage the risks, rewards, and ethical implications of AI technology and their own liability. Next, psychological factors that mediate how humans and AI technologies interact and the resulting impact on privacy are presented. The book concludes with a discussion of AI applications ranging from healthcare to warfare. These sections present the reader with real world situations and dilemmas that will impact stakeholders around the world. The chapter that follows introduces the reader to ethics and AI with an example that many people can try at home.

Chapter 2
What Is AI?

In this chapter we discuss the different definitions of Artificial Intelligence (AI). We then discuss how machines learn and how a robot works in general. Finally we discuss the limitations of AI and the influence the media has on our preconceptions of AI.

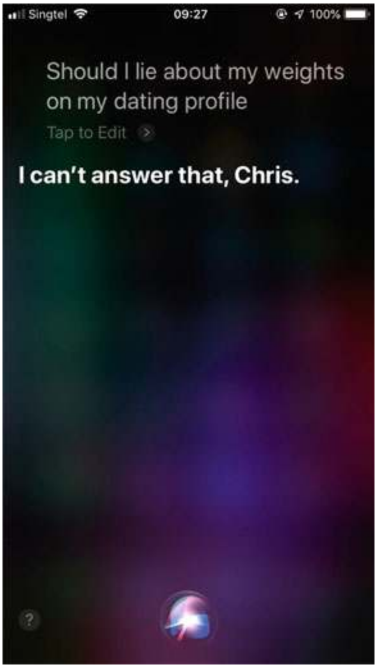
CHRIS: Siri, should I lie about my weight on my dating profile?
SIRI: I can't answer that, Chris.

Siri is not the only virtual assistant that will struggle to answer this question (see Fig. 2.1). Toma et al. (2008) showed that almost two thirds of people provide inaccurate information about their weight on dating profiles. Ignoring, for a moment, what motivates people to lie about their dating profiles, why is it so difficult, if not impossible, for digital assistants to answer this question?

To better understand this challenge it is necessary to look behind the scene and to see how this question is processed by Siri. First, the phone's microphone needs to translate the changes in air pressure (sounds) into a digital signal that can then be stored as data in the memory of the phone. Next, this data needs to be sent through the internet to a powerful computer in the cloud. This computer then tries to classify the sounds recorded into written words. Afterwards, an artificial intelligence (AI) system needs to extract the meaning of this combination of words. Notice that it even needs to be able to pick the right meaning for the homophone "lie". Chris does not want to lie down on his dating profile, he is wondering if he should put inaccurate information on it.

While the above steps are difficult and utilise several existing AI techniques, the next step is one of the hardest. Assuming Siri fully understands the meaning of Chris's question, what advice should Siri give? To give the correct advice, it would need to know what a person's weight means and how the term relates to their attractiveness. Siri needs to know that the success of dating depends heavily on both

C. Bartneck et al., *An Introduction to Ethics in Robotics and AI*,
SpringerBriefs in Ethics, https://doi.org/10.1007/978-3-030-51110-4_2



Should I lie about my weights on my dating profile

I can't answer that, Chris.

Fig. 2.1 Siri's response to a not so uncommon question

participants considering each other attractive—and that most people are motivated to date. Furthermore, Siri needs to know that online dating participants cannot verify the accuracy of information provided until they meet in person. Siri also needs to know that honesty is another attribute that influences attractiveness. While deceiving potential partners online might make Chris more attractive in the short run, it would have a negative effect once Chris meets his date face-to-face.

But this is not all. Siri also needs to know that most people provide inaccurate information on their online profiles and that a certain amount of dishonesty is not likely to impact Chris's long-term attractiveness with a partner. Siri should also be aware that women select only a small portion of online candidates for first dates and that making this first cut is essential for having any chance at all of convincing the potential partners of Chris's other endearing qualities.

There are many moral approaches that Siri could be designed to take. Siri could take a consequentialist approach. This is the idea that the value of an action depends on the consequences it has. The best known version of consequentialism is the classical utilitarianism of Jeremy Bentham and John Stuart Mill (Bentham 1996; Mill 1863). These philosophers would no doubt advise Siri to maximise happiness: not just Chris's happiness but also the happiness of his prospective date. So, on the consequentalist approach Siri might give Chris advice that would maximise his chances to not only to have many first dates, but maximise the chances for Chris to find true love.

Alternatively, Siri might be designed to take a deontological approach. A deontologist like Immanuel Kant might prioritise duty over happiness. Kant might advise Chris that lying is wrong. He has a duty not to lie so he should tell the truth about his weight, even if this would decrease his chances of getting a date.

A third approach Siri could take would be a virtue ethics approach. Virtue ethics tend to see morality in terms of character. Aristotle might advise Chris that his conduct has to exhibit virtues such as honesty.

Lastly, Siri needs to consider whether it should give a recommendation at all. Providing wrong advice might damage Siri's relationship to Chris and he might consider switching to another phone with another digital assistant. This may negatively impact Apple's sales and stock value.

This little example shows that questions that seem trivial on the surface might be very difficult for a machine to answer. Not only do these machines need the ability to process sensory data, they also need to be able to extract the correct meaning from it and then represent this meaning in a data structure that can be digitally stored. Next, the machine needs to be able to process the meaning and conclude with desirable actions. This whole process requires knowledge about the world, logical reasoning and skills to learn and adapt. Having these abilities may make the machine **autonomous**.

There are various definitions of "autonomy" and "autonomous" in AI, robotics and ethics. At its simplest, autonomous simply refers to the ability of a machine to operate for a period of time without a human operator. Exactly what that means differs from application to application. What is considered "autonomous" in a vehicle is different to what is considered "autonomous" in an weapon. In bioethics autonomy refers to the ability of humans to make up their own minds about what treatment to accept or refuse. In Kantian ethics autonomy refers to the ability of humans to decide what to do with their lives and what moral rules to live by. The reader should be aware that exactly what "autonomous" means is context-sensitive. Several meanings are presented in this book. The unifying underlying idea is self-rule (from the Greek words "auto" meaning self and "nomos" meaning rule).

On the first of these definitions, Siri is an autonomous agent that attempts to answer spoken questions. Some questions Siri tries to answer require more intelligence, meaning more background, reasoning ability and knowledge, than others. The chapter that follows define and describe the characteristics that make something artificially intelligent and an agent.

2.1 Introduction to AI

The field of artificial intelligence (AI) has evolved from humble beginnings to a field with global impact. The definition of AI and of what should and should not be included has changed over time. Experts in the field joke that AI is everything that computers cannot currently do. Although facetious on the surface, there is a sense that developing intelligent computers and robots means creating something that does not exist today. Artificial intelligence is a moving target.

Indeed, even the definition of AI itself is volatile and has changed over time. Kaplan and Haenlein define AI as “a system’s ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation” (Kaplan and Haenlein 2019). Poole and Mackworth (2010) define AI as “the field that studies the synthesis and analysis of computational agents that act intelligently.” An agent is something (or someone) that acts. An agent is intelligent when:

1. its actions are appropriate for its circumstances and its goals
2. it is flexible to changing environments and changing goals
3. it learns from experience, and
4. it makes appropriate choices given its perceptual and computational limitations.

Russell and Norvig define AI as “the study of [intelligent] agents that receive precepts from the environment and take action. Each such agent is implemented by a function that maps percepts to actions, and we cover different ways to represent these functions, such as production systems, reactive agents, logical planners, neural networks, and decision-theoretic systems” Russell and Norvig (2010, p. viii).

Russell and Norvig also identify four schools of thought for AI. Some researchers focus on creating machines that think like humans. Research within this school of thought seeks to reproduce, in some manner, the processes, representations, and results of human thinking on a machine. A second school focuses on creating machines that act like humans. It focuses on action, what the agent or robot actually does in the world, not its process for arriving at that action. A third school focuses on developing machines that act rationally. Rationality is closely related to optimality. These artificially intelligent systems are meant to always do the right thing or act in the correct manner. Finally, the fourth school is focused on developing machines that think rationally. The planning and/or decision-making that these machines will do is meant to be optimal. Optimal here is naturally relevant to some problems that the system is trying to solve.

We have provided three definitions. Perhaps the most basic element common to all of them is that AI involves the study, design and building of intelligent agents that can achieve goals. The choices an AI makes should be appropriate to its perceptual and cognitive limitations. If an AI is flexible and can learn from experience as well as sense, plan and act on the basis of its initial configuration, it might be said to be more intelligent than an AI that just has a set of rules that guides a fixed set of actions. However, there are some contexts in which you might not want the AI to learn new rules and behaviours, during the performance of a medical procedure, for example. Proponents of the various approaches tend to stress some of these elements more than others. For example, developers of expert systems see AI as a repository of expert knowledge that humans can consult, whereas developers of machine learning systems see AI as something that might discover new knowledge. As we shall see, each approach has strengths and weaknesses.

2.1.1 The Turing Test

In 1950 Alan Turing (see Fig. 2.2) suggested that it might be possible to determine if a machine is intelligent based on its ability to exhibit intelligent behaviour which is indistinguishable from an intelligent human's behaviour. Turing described a conversational agent that would be interviewed by a human. If the human was unable to determine whether or not the machine was a person then the machine would be viewed as having passed the test. Turing's argument has been both highly influential and also very controversial. For example, Turing does not specify how long the

Fig. 2.2 Alan Turing (1912–1954) (*Source* Jon Callas)

human would have to talk to the machine before making a decision. Still, the Turing Test marked an important attempt to avoid ill-defined vague terms such as "thinking" and instead define AI with respect to a testable task or activity.

2.1.2 Strong and Weak AI

John Searle later divided AI into two distinct camps. Weak AI is limited to a single, narrowly defined task. Most modern AI systems would be classified in this category. These systems are developed to handle a single problem, task or issue and are generally not capable of solving other problems, even related ones. In contrast to weak AI, Searle defines strong AI in the following way: "The appropriately programmed computer with the right inputs and outputs would thereby have a mind in exactly the same sense human beings have minds" (Searle 1980). In strong AI, Searle chooses to connect the achievement of AI with the representation of information in the human mind. While most AI researchers are not concerned with creating an intelligent agent that meets Searle's strong AI conditions, these researchers seek to eventually create machines for solving multiple problems which are not narrowly defined. Thus one of the goals of AI is to create autonomous systems that achieve some level of general intelligence. No AI system has yet achieved general intelligence.

2.1.3 Types of AI Systems

There are many different types of AI systems. We will briefly describe just a few. Knowledge representation is an important AI problem that tries to deal with how information should be represented in order for a computer to organise and use this information. In the 1960s, expert systems were introduced as knowledge systems that can be used to answer questions or solve narrowly defined problems in a particular domain. They often have embedded rules that capture knowledge of a human expert. Mortgage loan advisor programs, for example, have long been used by lenders to evaluate the credit worthiness of an applicant. Another general type of AI system are planning systems. Planning systems attempt to generate and organise a series of actions which may be conditioned on the state of the world and unknown uncertainties. The Hubble telescope, for example, utilised an AI planning system called SPIKE.

Computer vision is a subfield of AI which focuses on the challenge of converting data from a camera into knowledge representations. Object recognition is a common task often undertaken by computer vision researchers. Machine learning focuses on developing algorithms the allow a computer to use experience to improve its performance on some well-defined task. Machine learning is described in greater detail in the sections below.

AI currently works best in constrained environments, but has trouble with open worlds, poorly defined problems, and abstractions. Constrained environments include simulated environments and environments in which prior data accurately reflects future challenges. The real world, however, is open in the sense that new challenges arise constantly. Humans use solutions to prior related problems to solve new problems. AI systems have limited ability to reason analogically from one situation to another and thus tend to have to learn new solutions even for closely related problems. In general, they lack the ability to reason abstractly about problems and to use common sense to generate solutions to poorly defined problems.

2.2 What Is Machine Learning?

Machine learning is a sub-field of AI focused on the creation of algorithms that use experience with respect to a class of tasks and feedback in the form of a performance measure to improve their performance on that task. Contemporary machine learning is a sprawling, rapidly changing field. Typically machine learning is sub-categorised into three types of learning.

Supervised learning centres on methods such as regression and classification. To solve a classification problem experiences in the form of data are labelled with respect to some target categorisation. The labelling process is typically accomplished by enlisting the effort of humans to examine each piece of data and to label the data. For supervised learning classification problems performance is measured by calculating the true positive rate (the ratio of the true positives over all positives, correctly labelled or not) and the false positive rate (the ratio of false positives over all negatively classified data, correctly and incorrectly labelled). The result of this machine learning process is called a *classifier*. A classifier is software that can automatically predict the label of a new piece of data. A machine learning classifier that categorises labelled data with a true positive rate of 100% and a false positive rate of 0% is a perfect classifier. The supervised learning process then is the process by which unlabelled data is fed to a developing classifier and, over the course of working through some training data, the classifier's performance improves. Testing the classifier requires the use of a second label data-set called the test data set. In practice, often one overall data-set is carved into a training and test set on which the classifier is then trained and tested. The testing and training process may be time-consuming, but once a classifier is created it can be used to quickly categorise incoming data.

Unsupervised learning is more focused on understanding data patterns and relations than on prediction. It involves methods such as principal components analysis and clustering. These are often used as exploratory precursors to supervised learning methods.

Reinforcement learning is a third type of machine learning. Reinforcement learning does not focus on the labelling of data, but rather attempts to use feedback in the form of a reinforcement function to label states of the world as more or less desirable with respect to some goal. Consider, for example, a robot attempting to move from one location to another. If the robot's sensors provide feedback telling it its distance from a goal location, then the reinforcement function is simply a reflection of the sensor's readings. As the robot moves through the world it arrives at different locations which can be described as states of the world. Some world states are more rewarding than others. Being close to the goal location is more desirable than being further away or behind an obstacle. Reinforcement learning learns a policy, which is a mapping from the robot's action to expected rewards. Hence, the policy tells the system how to act in order to achieve the reward.

2.3 What Is a Robot?

Typically, an artificially intelligent agent is software that operates online or in a simulated world, often generating perceptions and/or acting within this artificial world. A robot, on the other hand, is situated in the real world, meaning that its existence and operation occur in the real world. Robots are also embodied, meaning that they have a physical body. The process of a robot making intelligent decisions is often described as "sense-plan-act" meaning that the robot must first sense the environment, plan what to do, and then act in the world.

2.3.1 Sense-Plan-Act

A robot's embodiment offers some advantages in that its experiences tend to be with real objects, but it also poses a number of challenges. Sensing in the real world is extremely challenging. Sensors such as cameras, laser scanners, and sonar all have limitations. Cameras, for example, suffer from colour shifts whenever the amount of light changes. Laser scanners have difficulty perceiving transparent objects. Converting sensor data into a usable representation is challenging and can depend on the nature and limitations of the sensor. Humans use a wide array of integrated sensors to generate perceptions. Moreover, the number of these sensors is (at least currently) much higher than the number of sensors of any robot. The vast amount of sensors available to a human is advantageous in terms of uncertainty reduction of perception. Humans also use a number different brain structures to encode information, to perform experience-based learning, and to relate this learning to other knowledge and experiences. Machines typically cannot achieve this type of learning.

Planning is the process by which the robot makes use of its perceptions and knowledge to decide what to do next. Typically, robot planning includes some type of goal that the robot is attempting to achieve. Uncertainty about the world must be dealt with at the planning stage. Moreover, any background or historical knowledge that the system has can be applied at this stage.

Finally, the robot acts in the world. The robot must use knowledge about its own embodiment and body schema to determine how to move joints and actuators in a manner dictated by the plan. Moreover, once the robot has acted it may need to then provide information to the sensing process in order to guide what the robot should look for next.

It should be understood that AI agents and robots have no innate knowledge about the world. Coming off the factory production line a robot or AI is a genuine "blank slate" or to be more exact an unformatted drive. Babies, on the other hand, enter the world "pre-programmed" so to speak with a variety of innate abilities and knowledge. For example, at birth babies can recognise their mother's voice. In contrast, AI agents know nothing about the world that they have not been explicitly programmed to know. Also in contrast to humans, machines have limited ability to generate knowledge from perception. The process of generating knowledge from information requires that the AI system creates meaningful representations of the knowledge. As mentioned above, a representation is a way of structuring information in order to make it meaningful. A great deal of research and debate has focused on the value of different types of representations. Early in the development of AI, symbolic representations predominated. A symbolic representation uses symbols, typically words, as the underlying representation for an object in the world. For example, the representation of the object apple would be little more than "Apple." Symbolic representations have the value of being understandable to humans but are otherwise very limiting because they have no precise connection to the robot's or the agent's sensors. Non-symbolic representations, on the other hand, tend not to be easily understood, but tend to relate better to a machine's sensors.

2.3.2 System Integration. Necessary but Difficult

In reality, to develop a working system capable of achieving real goals in the real world, a vast array of different systems, programmes and processes must be integrated to work together. System integration is often one of the hardest parts of building a working robotic system. System integrators must deal with the fact that different information is being generated by different sensors at different times. The different sensors each have unique limitations, uncertainties, and failure modes, and the actuators may fail to work in the real world. For all of these reasons, creating artificially intelligent agents and robots is extremely challenging and fraught with difficulties.

2.4 What Is Hard for AI

The sections above have hinted at why AI is hard. It should also be mentioned that not all software is AI. For example, simple sorting and search algorithms are not considered intelligent. Moreover, a lot of non-AI is smart. For example, control algorithms and optimisation software can handle everything from airline reservation systems to the management of nuclear power plants. But they only take well-defined actions within strictly defined limits. In this section, we focus on some of the major challenges that make AI so difficult. The limitations of sensors and the resulting lack of perception have already been highlighted.

AI systems are rarely capable of generalising across learned concepts. Although a classifier may be trained on very related problems, typically classifier performance drops substantially when the data is generated from other sources or in other ways. For example, face recognition classifiers may obtain excellent results when faces are viewed straight on, but performance drops quickly as the view of the face changes to, say profile. Considered another way, AI systems lack robustness when dealing with a changing, dynamic, and unpredictable world. As mentioned, AI systems lack common sense. Put another way, AI systems lack the enormous amount of experience and interactions with the world that constitute the knowledge that is typically called common sense. Not having this large body of experience makes even the most mundane task difficult for a robot to achieve. Moreover, lack of experience in the world makes communicating with a human and understanding a human's directions difficult. This idea is typically described as common ground.

Although a number of software systems have claimed to have passed the Turing test, these claims have been disputed. No AI system has yet achieved strong AI, but some may have achieved weak AI based on their performance on a narrow, well-defined task (like beating a grandmaster in chess or Go, or experienced players in Poker). Even if an AI agent is agreed to have passed the Turing test, it is not clear whether the passing of the test is a necessary and sufficient condition for intelligence.

AI has been subject to many hype cycles. Often even minor advancements have been hailed as major breakthroughs with predictions of soon to come autonomous intelligent products. These advancements should be considered with respect to the narrowness of the problem attempted. For example, early types of autonomous cars capable of driving thousands of miles at a time (under certain conditions) were already being developed in the 1980s in the US and Germany. It took, however, another 30+ years for these systems to just begin to be introduced in non-research environments. Hence, predicting the speed of progression of AI is very difficult—and in this regard, most prophets have simply failed.

2.5 Science and Fiction of AI

Artificial Intelligence and robotics are frequent topics in popular culture. In 1968, the Stanley Kubrick classic "2001" featured the famous example of HAL, a spacecraft's intelligent control system which turns against its human passengers. The Terminator movies (since 1984) are based on the idea that a neural network built for military defense purposes gains self-awareness and, in order to protect itself from deactivation by its human creators, turns against them. The Steven Spielberg's movie "A.I." (2001), based on a short story by Brian Aldiss, explores the nature of an intelligent robotic boy (Aldiss 2001). In the movie “I, Robot” (2004), based on motives from a book by Isaac Asimov, intelligent robots originally meant to protect humans are turning into a menace. A more recent example is the TV show “Westworld” (since 2016) in which androids entertain human guests in a Western theme park. The guests are encouraged to live out their deepest fantasies and desires.

For most people, the information provided through these shows is their first exposure to robots. While these works of fiction draw a lot of attention to the field and inspire our imagination, they also set a framework of expectations that can inhibit the progress of the field. One common problem is that the computer systems or robots shown often exhibit levels of intelligence that are equivalent or even superior to that of humans or current systems. The media thereby contributes to setting very high expectations in the audience towards AI systems. When confronted with actual robots or AI systems, people are often disappointed and have to revise their expectations. Another issue is the frequent repetition of the “Frankenstein Complex” as defined by Isaac Asimov. In this trope, bands of robots or an AI system achieve consciousness and enslave or kill (all) humans. While history is full of examples of colonial powers exploiting indigenous populations, it does not logically follow that an AI system will repeat these steps. A truly intelligent system will (hopefully) have learned from humanity’s mistakes. Another common and rather paradoxical trope is the assumption that highly intelligent AI systems desire to become human. Often the script writers use the agent’s lack of emotions as a the missing piece of the puzzle that would make them truly human.

It is important to distinguish between science and fiction. The 2017 recommendation to the European Parliament to consider the establishment of electronic personalities (Delvaux 2017) has been criticised by many as a premature reflex to the depiction of robots in the media.[1] For example, granting the robot “Sophia” Saudi Arabian citizenship in October 2017 can in this respect be considered more as a successful public relations stunt (Reynolds 2018) than as a contribution to the field of AI or its ethical implications. Sophia’s dialogues are based on scripts and cannot therefore be considered intelligent. It does not learn nor is it able to adapt to unforeseen circumstances. Sophia’s presentation at the United Nation is an unconvincing demonstration of artificial intelligence. People do anthropomorphise robots and autonomous systems, but this does not automatically justify the granting of personhood or other forms of legal status. In the context of autonomous vehicles, it may become practical to consider such a car a legal entity, similar to how we consider an abstract company to be a legal person. But this choice would probably be motivated more out of legal practicality than out of existential necessity.

[1] http://www.robotics-openletter.eu/.

Discussion Questions:

- Explain the difference between weak and strong AI. Give examples from science fiction describing machines that could be categorised as displaying strong and weak AI.
- Given the description of supervised machine learning above, how might a classifier come to include societal biases? How might the removal of such biases impact classifier performance? Describe a situation in which stakeholders must balance the tradeoff between bias and performance.
- Consider the sense-plan-act paradigm described above. How might errors at one step of this process impact the other steps? Draw an informal graph of robot performance versus time.

Further Reading:

- Stuart J. Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Prentice Hall, Upper Saddle River, N.J, 3rd edition, 2010. ISBN 9780132071482. URL http://www.worldcat.org/oclc/688385283
- Ryszard S Michalski, Jaime G Carbonell, and Tom M Mitchell. Machine learning: *An artificial intelligence approach*. Springer Science & Business Media, 2013. ISBN 978-3662124079. URL http://www.worldcat.org/oclc/864590508
- Sidney Perkowitz. *Digital people: From bionic humans to androids*. Joseph Henry Press, 2004. ISBN 978-0309096195. URL http://www.worldcat.org/oclc/936950712.

Chapter 3
What Is Ethics?

This chapter introduces the theories that form the basis of the ethical review of robots and AI systems. We introduce the major approaches to moral theory (deontology, consequentialism and virtue ethics) and discuss the relation of ethics to law. Finally, we discuss how these theories might be implemented in a machine to enable it to make ethical decisions.

The terms "ethics" and "morality" are often taken as synonyms. Sometimes they are distinguished, however, in the sense that morality refers to a complex set of rules, values and norms that determine or are supposed to determine people's actions, whereas ethics refers to the theory of morality. It could also be said that ethics is concerned more with principles, general judgements and norms than with subjective or personal judgements and values.

Etymologically, the word ethics goes back to the ancient Greek "ethos". This originally referred to a place of dwelling, location, but also habit, custom, convention. It was Cicero who translated the Greek term into Latin with "mores" (ethos, customs), from which the modern concept of morality is derived (Cicero 44BC). The German philosopher Immanuel Kant (see Fig. 3.1) characterised ethics as dealing with the question "What should I do?" (Kant 1788). There are several schools of thought on ethics and we will introduce them in here in no particular order.

3.1 Descriptive Ethics

Most people, when thinking of ethics, have normative ethics in mind as described below. Like ethnology, moral psychology or experimental economics, descriptive ethics deals with the description and explanation of normative systems. For example,

C. Bartneck et al., *An Introduction to Ethics in Robotics and AI*,
SpringerBriefs in Ethics, https://doi.org/10.1007/978-3-030-51110-4_3

Fig. 3.1 Immanuel Kant (1724–1804) (*Source* Johann Gottlieb Becker)

experimental results exhibit certain features of moral intuitions of people: studies using the “ultimatum game” show that many people have certain intuitions about fairness and are willing to sacrifice profits for these intuitions (Güth et al. 1982). These empirical insights form the basis of descriptive ethics which in turn provides essential input for normative ethics. Normative evaluation of actions is not possible without descriptive elements of empirical insights. In recent years, “experimental ethics” has even been formed as a sub-discipline of its own (Lütge et al. 2014). For the remainder of the book, we use the term ‘ethics’ to mean ‘normative ethics.’

3.2 Normative Ethics

Ethics can be defined as the analysis of human actions from the perspective of “good” and “evil,” or of “morally correct” and “morally wrong.” If ethics categorises actions and norms as morally correct or wrong, one then speaks of normative or prescriptive ethics. An example of a norm is that the action of stealing is morally wrong. Normative ethics is usually not regarded as a matter of subjectivity, but of general validity. Stealing is wrong for everybody. Different types of normative ethics make judgements about actions on the basis of different considerations. The most important distinction usually made here is between two types of theories: deontological and consequentialist ethics.

3.2.1 Deontological Ethics

Deontological ethics is characterised by the fact that it evaluates the ethical correctness of actions on the basis of characteristics that affect the action itself. Such a feature, for example, may be the intention with which an action is performed or the compatibility with a particular formal principle. The consequences of an action may be considered in addition, but does not form the exclusive basis of the judgement. The term deontology or deontological ethics derives from the Greek “deon”, which essentially means duty or obligation. Deontology can thus be translated as duty ethics.

To give a practical example of deontological ethics, since the 2000s large and medium-sized companies have increasingly tried to project a social or environmentally friendly image through certain marketing and PR measures. Often, as part of these measures, companies donate sizeable sums to combat certain social ills, improve their environmental footprint, or work with NGOs to more effectively monitor the conditions of production among suppliers. Nevertheless, many citizens refuse to positively assess this commitment of companies as ethically genuine. The public discussion sometimes ridicules such programmes of Corporate Social Responsibility (CSR). Critics argue that in these cases companies are not really concerned with rectifying grievances, but only with polishing up their own image and ultimately maximising their bottom line, albeit in a more sophisticated way. Regardless of whether the CSR projects in question contribute to improving some of the (environmental or social) issues, critics are more concerned with the companies motivations than with their action or the results. The companies motivations being the key deontological element for this argument.

Kant is responsible for developing one of the most frequently cited deontological ethics. He argues that an action is only obligatory if it satisfies the "categorical imperative". There are many different wordings of the categorical imperative, which is best understood as a way of determining ethically permissible types of behaviour. The most frequently cited version states, "Act only according to that maxim you can at the same time will as a universal law without contradiction." (see Sect. 4.3.3).

3.2.2 Consequentialist Ethics

Consequentialism is another important ethical theory. Consequentialist theories determine the ethical correctness of an action or a norm solely on the basis of their (foreseeable) consequences. The difference between consequentialism and deontological ethics can be seen in the previously used example. From the perspective of consequentialism, the motives of a company to invest in CSR play no role. For this ethical evaluation of a company's CSR programme, the only decisive considerations relate to the impact on society, wildlife, nature or maybe social harmony. As long as a CSR programme promotes certain values or, more generally, helps to solve certain social problems, the program can be considered ethical. This also applies if a particular CSR programme was merely motivated by the desire to improve the image of a company or increase sales.

3.2.3 Virtue Ethics

The concept of virtue ethics mainly goes back to the Greek philosophers Plato (see Fig. 3.2), who developed the concept of the cardinal virtues (wisdom, justice, fortitude, and temperance), and Aristotle, who expanded the catalogue into eleven moral virtues and even added intellectual virtues (like Sophia = theoretical wisdom). The classical view on virtues held that acting on their basis was equally good for the person acting and for the persons affected by their actions. Whether this is still the case in modern differentiated societies is controversial.

3.3 Meta-ethics

If ethics can be regarded as the theory of morality, meta-ethics is the theory of (normative) ethics. Meta-ethics is concerned, in particular, with matters of existence (ontology), meaning (semantics) and knowledge (epistemology). Moral ontology is an account of what features of the world have moral significance or worth. Moral

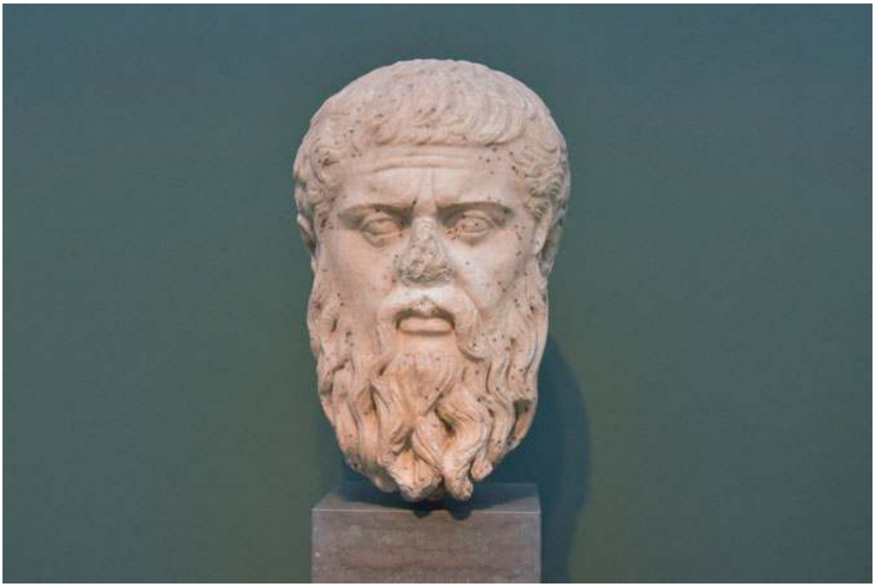


Fig. 3.2 Plato (*Source* Richard Mortel)

semantics is an account of the meaning of moral terms such as right, wrong, good, bad and ought to name the most prominent. Moral epistemology is an account of how we can know moral truth.

3.4 Applied Ethics

Normative and meta-ethics are usually distinguished from applied ethics. Applied ethics refers to more concrete fields where ethical judgements are made, for example in the areas of medicine (medical ethics), biotechnology (bioethics) or business (business ethics). In this sense, general normative considerations can be distinguished from more applied ones. However, the relation between the two should not be seen as unidirectional, in the sense that general ("armchair") considerations come first and are later applied to the real world. Rather, the direction can be both ways, with special conditions of an area in question bearing on general questions of ethics. For example, the general ethical principle of solidarity might mean different things under different circumstances. In a small group, it might imply directly sharing certain goods with your friends and family. In a larger group or in an entire society, however, it might imply quite different measures, such as competing fairly with each other.

3.5 Relationship Between Ethics and Law

Often, ethics and law are seen as being clearly distinct from each other, sometimes even as opposites, in the sense, for example, that ethics starts where the law ends. Persons or companies would then have legal duties and ethical duties, which have little relationship with each other. However, such a view can be challenged, in several ways. First, legal rules often have an ethical side, too. For example, legal norms that make environmental pollution illegal still remain ethical norms, too. Much of the legal framework of a society (like anti-trust laws) has a great ethical importance for a society. Second, ethics can (and has) to some extent become a kind of "soft law", in the sense that companies need to follow certain ethical standards even if the law in a particular country does not strictly require it. For fear of damaging their reputation, or decreasing the value of their stock, for example, companies are, in many cases, adhering to ethical rules which for them have nearly the same consequences and impact as legal rules ("hard law"). At times the specific ethical process of a business is even used as a unique sales argument, such as companies selling "fair trade" coffee instead or just plain legal coffee.

3.6 Machine Ethics

Machine ethics attempts to answer the question: what would it take to build an ethical AI that could make moral decisions? The main difference between humans making moral decisions and machines making moral decisions is that machines do not have "phenomenology" or "feelings" in the same way as humans do (Moor 2006). They do not have "moral intuition" or "acculturation" either. Machines can process data that represents feelings (Sloman and Croucher 1981), however, no one, as yet, supposes that computers can actually feel and be conscious like people. Life-like robots have been developed (e.g. Hanson Robotics Sophia—see Fig. 3.3) but these robots do not possess phenomenal consciousness or actual feelings of pleasure or pain. In fact, many argue that the robot Sophia represents more of a corporate publicity stunt than a technological achievement and, as such, represents how the mystique around robots and artificial intelligence can be harnessed for attention. In this section we discuss how to design an AI that is capable of making moral decisions. Technical and philosophical elements are presented. Yet we should note that the goal of creating machines that make moral decisions is not without detractors (van Wynsberghe and Robbins 2019). Van Wynsberghe and Robbins note that outside of intellectual curiosity, roboticists have generally failed to present strong reasons for developing moral robots.

The philosophical element is a detailed moral theory. It will provide us with an account of what features of the world have moral significance and a decision procedure that enables us to decide what acts are right and wrong in a given situation.

Fig. 3.3 The Sophia robot (*Source* Hanson Robotics)

Such a decision procedure will be informed by a theory as to what acts are right and wrong in general.

For philosophical convenience, we assume our AI is embedded in a humanoid robot and can act in much the same way as a human. This is a large assumption but for the moment we are embarking on a philosophical thought experiment rather than an engineering project.

3.6.1 Machine Ethics Examples

The process for an ethical AI embedded in a robot starts with sensor input. We assume sensor input can be converted into symbols and that these symbols are input into a moral cognition portion of the robot's control system. The moral cognition system must determine how the robot should act. We use the term symbol grounding to refer to the conversion of raw sensor data to symbols. Symbols are used to represent objects and events, properties of objects and events, and relations between objects and events.

Reasoning in an AI typically involves the use of logic. Logic is truth-preserving inference. The most famous example of logical deduction comes from Aristotle. From two premises, "Socrates is a man" and "all men are mortal" the conclusion

"Socrates is mortal" can be proved. With the right logical rules, the premises we may need to deduce action will be based on symbols sensed in the environment by the robot.

To illustrate, we can assume that our robot is tasked with issuing tickets to speeding cars. We can also assume also that the minimal input the system needs to issue a ticket is the symbol representing the vehicle (e.g. the license plate number) and a symbol representing whether or not the vehicle was speeding (Speeding or NOT Speeding). A logical rule of of inference can be stated as "If driver X is speeding then the robot U is obligated to issue a ticket to driver X." In much the same was as we can deduce "Socrates is a mortal" from two premises, we can derive a conclusion such as "the robot U is obligated to issue ticket" from the rule of inference above and a statement like "the driver of car X is speeding."

Now consider a more difficult problem for the machine that is still morally obvious to a human. Imagine a robot is walking to the post office to post a letter. It walks along a path by a stream. Suddenly a toddler chases a duck which hops into the stream. The toddler slips and falls into the water which is one metre deep. The toddler is in imminent danger of drowning. The robot is waterproof. Should it enter the water and rescue the toddler or should it post the letter? This question is morally obvious to a human but a robot does not have empathy and feelings of urgency and emergency. Hence, it needs rules to make the decision. To solve the toddler problem the robot must have an understanding of causation. If the toddler remains in the water he or she will drown but the letter can be posted at any time. If the robot rescues the toddler, it will not drown, but it may miss an opportunity to post the letter. As a result the letter will arrive a day late.

How does the robot determine what to do? It needs to represent the value of a saved life compared to the value of a one day delay in the arrival of a posted letter at its destination. In deontological terms the robot has two duties, to save the toddler and to post the letter. One has to be acted on first and one deferred. To resolve the clash between duties we need a scale on which the value of the consequences of the two actions can be compared. Such a scale would value saving the toddler's life over promptly posting a letter. One can assign a utility (a number) to each outcome. The robot can then resolve the clash of duties by calculating these utilities. Thus to solve the toddler versus delayed letter problem we compare the two numbers. Clearly the value of a saved life is orders of magnitude larger than a delayed letter. So, the duty to post the letter can yield to the duty to save the toddler.

Now suppose that the value we give to the living toddler was $+1{,}000{,}000$ and the value we gave to an on time letter was $+1$. Clearly there are orders of magnitude of difference between the value of the toddler and a promptly posted letter. Now consider a scenario in which the robot is travelling to the post office driving a truck with a million and one letters each valued $+1$. The robot sees the toddler fall into the stream and using the same logic as was used in the previous example determines not to stop and help the toddler. The moral arithmetic in this case is 1,000,001 to post the letters versus 1,000,000 to save the toddler. It is a narrow call but posting the letters wins by one point. To implement deontology in machines one needs a way to resolve clashes between duties. If you take a naive consequentialist approach and just assign simple utilities to outcomes you run the risk of running into counter-intuitive dilemmas such as this illustrative example.

3.6.2 Moral Diversity and Testing

One of the main challenges for machine ethics is the lack of agreement as to the nature of a correct moral theory. This is a fundamental problem for machine ethics. How do we implement moral competence in AIs and robots if we have no moral theory to inform our design?

One could design ethical test cases that an AI has to pass. Ideally we would create many test cases. Moral competence can be defined with respect to the ability to pass these test cases. In theory, as an agent or robot goes through iterative cycles of responding to new test cases its moral competence would expand. In so doing, one might gain insights into moral theory.

Testing and even certifying if an AI is fair and ethical is currently an important area of research. The Institute of Electrical and Electronics Engineers (IEEE) announced a Standards Project that addresses algorithmic bias considerations in 2017. Toolkits have been created that help developers to test if their software does have a bias, such as the AI Fairness 360 Open Source Toolkit,[1] audit-AI.[2] Some companies offer services to test the bias of algorithms, such as O'Neil Risk Consulting and Algorithmic Auditing,[3] and even big companies like Facebook are working on Fairness Flow, a tool to test biases. Keeping in mind that this is an area of ongoing inquiry, it should be noted that some researchers are pessimistic about the prospects for machine morality. Moreover, a number of research groups have developed or are developing codes of ethics for robotics engineers (Ingram et al. 2010) and the human-robot interaction profession (Riek and Howard 2014).

Discussion Questions:

- Explain the difference between normative ethics and meta-ethics.
- How would you judge the behaviour of a manager of an AI company who improves the fairness of their algorithms in order to increase the profit of their company? Discuss both from a deontological and from a consequentialist point of view.
- Do you think AI can become ethical? Can ethics be programmed into a machine? Discuss.

[1] https://aif360.mybluemix.net/.

[2] https://github.com/pymetrics/audit-ai.

[3] http://www.oneilrisk.com/.

Further Reading:

- Simon Blackburn. Being good: A short introduction to ethics. OUP Oxford, 2002. ISBN 978-0192853776. URL http://www.worldcat.org/oclc/945382272
- Stephen Darwall. Philosophical Ethics: An Historical and Contemporary Introduction. Routledge, 1997. ISBN 978-0813378602. URL http://www.worldcat.org/oclc/1082497213
- John Mackie. Ethics: Inventing right and wrong. Penguin UK, 1991. ISBN 978-0140135589. URL http://www.worldcat.org/oclc/846989284
- Christoph Luetge. Handbook of the philosophical foundations of business ethics. 2013. URL https://doi.org/10.1007/978-94-007-1494-6.

Chapter 4
Trust and Fairness in AI Systems

This chapter discusses the role that trust and fairness play in the acceptance of AI systems. We then relate trust and fairness in AI systems to five principles: Non-maleficence, Beneficence, Autonomy, Justice and Explicability.

There are many contexts in which one can use the word trust. For example, one might ask a person if they trust their spouse, if they trust a company such as Facebook or Google with their data, if they trust the government with their health data, or if they trust their housekeeper or their babysitter.

In human contexts, these questions all involve some kind of vulnerability to harm. Facebook and Google might sell data to some disreputable firm seeking to use it for manipulative purposes. People might believe that the government will hand over their health data to insurers which might raise their premiums. A housekeeper might steal. A babysitter might neglect the children in his or her care. In all of these cases, the person trusting has a vulnerability to some kind of harm or damage. If we do not trust people, we will be on our guard. We will not entrust them with anything important. We might even seek to avoid them completely if we can. Furthermore, uncertainty is a prerequisite for trust because vulnerability is diminished if one knows the outcome upfront.

Lee and See (2004) defined trust as "the attitude that an agent will help achieve an individual's goals in a situation by uncertainty and vulnerability." In the context of robots and AI systems, the question of trust is similar but differs in some respects. When a human trusts another human, the trustor knows that they are making themselves vulnerable to the actions of the other person. When a person trusts an AI agent, it is unclear if the machine is making it's own decision or following some predefined scripted behaviour. Furthermore, people have experiences with other people and have certain expectations about the behaviours, norms and values. We may trust a pilot to safely land an aircraft. When a user is interacting with a robot or an AI system, these experiences may become useless or even misleading. We cannot be

C. Bartneck et al., *An Introduction to Ethics in Robotics and AI*,
SpringerBriefs in Ethics, https://doi.org/10.1007/978-3-030-51110-4_4

certain that a system will act in our best interest. The rest of this chapter discusses trust and fairness in the context of humans trusting machines such as AIs and robots.

4.1 User Acceptance and Trust

Trust is critical to user acceptance. People will avoid using systems they do not trust. Obviously, businesses that make products that are not trusted will not see their products being used—and will thus struggle in the marketplace. For example, companies in the U.S. are obliged to comply with the Patriot Act that empowers the government to access data stored on cloud computers. Customers from Europe might feel uncomfortable about granting the U.S. government such a privilege. After the European Court of Justice overturned the long-standing US-EU Safe Harbor agreements in 2015, several cloud storage companies opened data centers in Europe to regain the trust of their European customers.

Culture influences the trust that people place in AI systems and robots. Haring et al. (2014a) shows that these cultural differences impact how people view robots. Cross-cultural studies also demonstrate differences in positivity (Haring et al. 2014b) which in turn impacts trust. Cultural factors may also influence the extent to which people follow a robot's recommendations (Wang et al. 2010).

In the case of machines, the question of trust can be broken up into "functional" and "ethical" elements.

4.2 Functional Elements of Trust

According to Lee and See (2004), users calibrate their trust in a machine based on a variety of factors including the system's reliability. Hancock et al. (2011) looked at a variety of factors influencing trust in automation. These included performance-based factors such as a system's dependability, false alarm rate, transparency, and task complexity. This work showed that performance and reliability are the dominant factors determining if a person trusts a machine.

People will be wary and suspicious of machines that are potentially harmful or dangerous. The possibility of loss of life or injury tends to cause people to reconsider whether or not to trust a machine.

4.3 Ethical Principles for Trustworthy and Fair AI

The European ethical principles for AI, presented by the AI4People group in 2018 (Floridi et al. 2018), suggest five principles for ethics of AI, which can be tied to trust and fairness in the following way.

4.3.1 Non-maleficence

The principle of Non-maleficence states that AI shall not harm people. AI systems that harm people (and animals or property) are a risk, both for individuals as well as for companies (cf. Chap. 6). In terms of trust and fairness, this extends to issues such as bullying and hate speech which are salient examples of the principle of non-maleficence being violated online.

4.3.1.1 Social Media Bullying and Harassment

The cases of school children being bullied by classmates or employees bullied at workplace are numerous. There are cases where this has even lead to suicide. While such bullying certainly existed prior to the advent of digital technologies, it has, through social networks, acquired a new level of concern, as indicated by surveys.[1] The risk of being exposed to large groups or to the public in general has risen dramatically. Consequently, in a number of countries, laws have been passed against bullying in digital media (Sweden was the first country to do so in 1993).

4.3.1.2 Hate Speech

Hate speech has in recent years gained considerable media coverage. Hate speech is a type of speech that is not just specifically directed against a particular person, but can also be directed against groups or entire parts of a population. It might, for example, attack groups on the basis of their ethnic origin, sexual orientation, religion, gender, identity, disability and others. Already the International Covenant on Civil and Political Rights", in force since 1976, includes a statement according to which "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law.

Laws against hate speech have since then been passed in a number of countries. For example, Belgium passed a law in 1981 which states that incitements to discrimination, hatred of violence against persons or groups of race, colour, origin or national or ethnic dissent are illegal and subject to punishment according to the Belgian penal code. Other countries like Canada, New Zealand or the United Kingdom have similar laws (which might however vary in detail and punishment).

A new level of hate speech laws was however reached in 2017 in Germany, when a bill was passed by the German Bundestag which specifically criminalises hate speech on social media. The law also insists that social networks (like Facebook or others) may be fined very large sums of up to 50 million EUR—in case they do not actively seek and successfully remove certain content within a week.

The passing of this law was controversial, with a number of German, but also many international commentators, stating that such a law is very far-reaching and

[1] https://www.theguardian.com/uk-news/2017/aug/14/half-uk-girls-bullied-social-media-survey.

will have a number of non-intended and counterproductive consequences. Since then, social networks like Twitter or Facebook have taken many efforts to comply with this new law. And while they have certainly succeeded in removing quite a lot of illegal content (as the European Commission acknowledged in 2019,[2] there have also been many issues with content being removed either by accident or due to over-interpretation of certain statements. Appeal of removal decisions is also important and social networks are starting to implement this.

The balance between hate speech and freedom of speech is being decided within and across nations. Social media has infused these debates with elements of cultural and national perspectives. Large companies such as Facebook and Google are being forced to edit their content. Some may view this as a necessary means for combating racism others may view this as in attack on freedom of speech.

4.3.2 Beneficence

The principle of Beneficence states that AI shall do people good. This is a general principle from bioethics according to which the benefits from a treatment must outweigh the potential harms (Zalta 2003). An AI system needs to consider ethical rules to become trustworthy and fair which will enable it to make life better. Examples of ways in which an AI system may demonstrate beneficence with respect to societal problems are:

- the reduction of fatalities and accidents by autonomous vehicles;
- providing robotic support and care for the ageing society (see Sect. 9.2)
- the use of telemedicine in remote areas;
- smart grid based sustainability improvements;
- improvements in terms of biodiversity, e.g., by AI applications to preserve endangered species
- the use of robots and AI in education, for example by using AI for individualised learning or supporting students outside the classroom.

4.3.3 Autonomy

The principle of Autonomy states that AI shall respect people's goals and wishes. To be sure, autonomy has several meanings, and we refer to some of them in Chaps. 10 and 11. Traditionally in AI and robotics, the term autonomy refers to an AI system's or robot's ability to operate without human intervention. In this section, however, we focus on the ethical principle of autonomy. In the context of bioethics, it usually refers to patients having the right to decide for themselves whether or not to undergo a

[2] http://fortune.com/2019/02/04/facebook-twitter-google-eu-hate-speech/.

treatment (Zalta 2003). This entails giving patients the right to refuse life-saving procedures and to decide not to take risk-reducing medications. For example, instances are known in which people died as a result of refusing blood transfusion for religious reasons. Generally, the courts have found that parents do not have a right to impose their views, such as refusing a blood transfusion, on their children (Woolley 2005). However, once a child becomes an adult, they can refuse treatment.

More generally, autonomy refers to the ability of a person to make decisions. People can decide whether or not they want to take risks to earn more money or have more fun. Sherpas carrying packs for mountaineers climbing Everest can earn five times more money than they can working in the fields of Nepal. However, they run the risk of being caught in an avalanche that can injure or kill them. People should be permitted to assume risks but they should know what risks they are assuming. An AI's respect of human autonomy would support permitting a degree of human self-harm and permitting humans to assume risks that might lead to harm. After all, humans might like to go rock-climbing or ride motorcycles.

There are ethical limits to autonomy. Suppose a child says to a robot, "pull my sister's hair, she's been mean to me and I hate her!" In this case, should the robot obey the child? If an adult says, "smack my daughter, she's been naughty" should the robot obey? Generally speaking, systems should not help people pursue illegal or immoral goals. Moreover, systems should not be used to perform illegal or immoral acts. There is no strong case for systems allowing users to employ the systems to harm others unless there is good reason. One should also be very careful when considering to design systems that permit machines to harm humans that ask to be harmed. For example, one might decline to develop an euthanasia robot. There is the risk such a robot might not sense mental illness. Also, one might hesitate to delegate such grave decisions to machines at all.

However, there are some cases in which robots are designed to use force against humans. The police may need to harm people in some cases such as apprehending violent offenders. For example, in 2016, the Dallas Police Chief ordered the use of a tele-operated robot to carry an explosive to a perpetrator who had shot 10 officers, killing 5 and wounding 5 (Thielmann 2016). The explosion killed the offender. Humans were in control of this particular robot. In the military domain, robots capable of violent action against humans have existed for some time. However, apart from lawful violence carried out by the state, there are relatively few cases where people are comfortable with robots and AIs being designed to harm humans.

Moral philosophy is directly linked to autonomy. If a person does not have autonomy or free will, then it can be argued that this person does not have moral responsibility either. From the deontological point of view the connection between autonomy and ethics can best be demonstrated by Immanuel Kant's moral theory (Kant 1785) which consists of three main formulations:

1. Act only according to that maxim you can at the same time will as a universal law without contradiction.

2. Act in such a way that you treat humanity, whether in your own person or in the person of any other, never merely as a means to an end, but always at the same time as an end.
3. The third practical principle follows from the first two as the ultimate condition of their harmony with practical reason: the idea of the will of every rational being as a universally legislating will.

The formula of universal law tells us that to morally justify an action, one must be able to universalise the moral rule or "maxim." One way to think about this is to pose the question: what if everybody followed this rule? What would the world look like then?

The principle of humanity tells us we should respect the ends (goals, wishes) of other persons. We should not treat other human beings as "mere means" alone. Rather we must consider the ends they have as well as our own.

Kant also formulated his moral philosophy in the so-called "Kingdom of Ends" version of the Imperative which states that: "Act according to maxims of a universally legislating member of a merely possible kingdom of ends." The principle of autonomy and the Kingdom of Ends formulation tell us that we must walk our own moral talk. That is, we are bound to obey the moral rules we expect others to follow. As autonomous beings, Kant holds, we are obliged to rationally consider our moral standards. Simply following the law is not good enough. It follows that an artificial intelligent system must have autonomy in the Kantian sense to be able to act in an ethical way.

Many writers bundle moral responsibility and moral decision making together in their definitions of what an "ethical agent" is. Some separate these two notions holding that an AI can make moral decisions without being responsible for such decisions. (Welsh 2018). On Kant's conception, a system that is programmed to simply follow rules such as Asimov's Three Laws of Robotics would not be considered an ethical agent.

Isaac Asimov (2 January 1920–6 April 1992) proposed three rules of robotics that would safeguard humanity from malevolent robots.

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 given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

While his work is highly visible in the public media, it has been criticised by philosophers. Asimov eventually added a zeroth law:

0. A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

4.3.4 Justice

The principle of justice states that AI shall act in a just and unbiased way. Justice is often illustrated by a statue of the Roman Goddess Justitia. Frequently, she is depicted with a sword, scales and a blindfold (see Fig. 4.1). The blindfold represents impartiality. The scales represent the weighing of evidence. The sword represents punishment.

Defining "justice" at the human level is a substantial challenge for AI. The main problem for AI is that moral theory is vigorously contested. There are a variety of "sects and schools" that have been engaged in "vigorous warfare" with each other since the "young Socrates debated the old Protagoras" as Mill (1863) puts it. Polling done by Bourget and Chalmers (2014) shows that none of the major schools of ethical theory enjoy firm majority support. Around a quarter of philosophers "accept" or "lean towards" either deontology or consequentialism. Around a third accept or lean towards virtue ethics. Thus generally defining "justice" or "ethics" in terms of what machines can process is hard. There is no agreement on moral theory.

Humans, by contrast, come with "moral intuition" and have learned some things about right and wrong over a number of years. It is often said by some that human

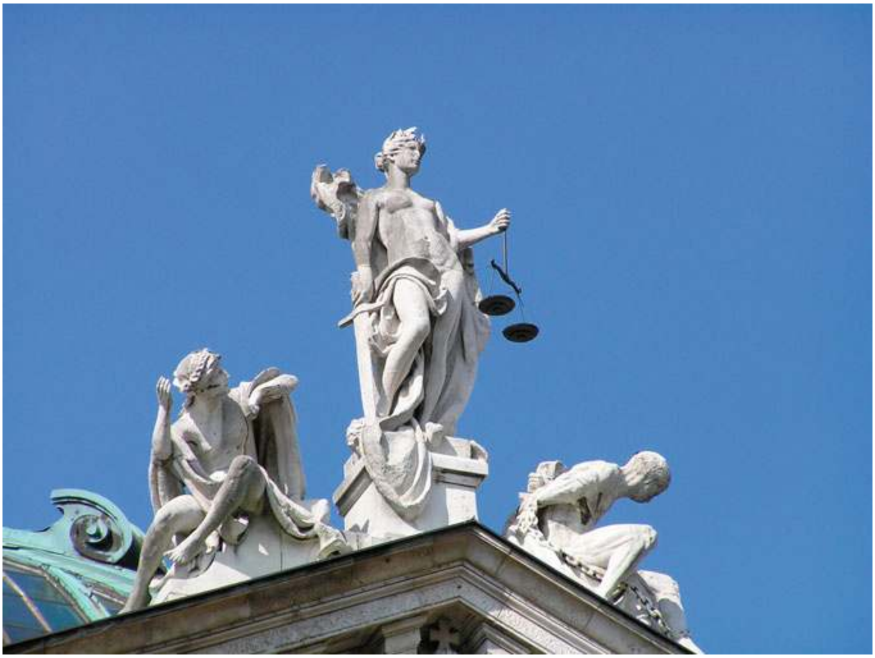


Fig. 4.1 Justitia (*Source* Waugsberg)

moral intuition is a “black box” of “inscrutable” biological code. We do not fully understand how humans make moral decisions. We do not even understand how human brains store information. However, while there is little agreement on moral theory, when it comes to moral practice, there are many actions that are generally considered to be “right” and many that are considered to be “wrong.” While ethical controversies rage on topics such as abortion, euthanasia, civil disobedience and capital punishment, there are many moral problems that are far less difficult.

If the scope of application is reduced and the information on which the moral decisions are made can be obtained, then it is possible for AI to make some very specific moral decisions. Often, there are clear regulations and laws that in very specific applications can be taken as normatively definitive. In practice, AI has already led to a number of implemented applications that may have moral and ethical implications.

4.3.4.1 Determining Creditworthiness

Banks and other credit institutions are already, in a number of countries, using AI systems to pre-sort credit applications on the basis of the data available about the applicant. This certainly has a number of advantages, one of which is to be able to come to a decision more quickly and on the basis of more information, making it more rational in theory. However, this may also entail disadvantages, in particular leading to certain biases. For example, the personal information of a credit applicant will in most cases contain information about their neighbourhood. On that basis, and making use of publicly available (or privately collected) further data, a systematic bias against persons from certain residential areas may occur.

An in-depth examination of racial bias related to the use of algorithms for high-risk care management demonstrates many of the challenges and issues associated with the merged concepts of justice, algorithmic fairness, and accuracy (Obermeyer et al. 2019). Researchers examined the inputs, outputs, and objective function to algorithm used to identify patients as high risk of needing acute care and thereby influenced the treatment of millions of Americans. Even though the algorithm specifically excludes race from consideration, the system reasonably uses information about healthcare costs to predict healthcare need. The system uses machine learning to create a model to predict future healthcare costs. It assumes that those individuals that will have the highest healthcare costs will be the same individuals that need the most healthcare, a very reasonable assumption. Yet, this assumption introduces disparities that end up correlating to race. For example, poor patients face greater challenge accessing healthcare because they may lack access to transportation, childcare, or have competing work related demands. They conclude that the central issue is problem formulation: the challenge of developing precise computational algorithms that operate on amorphous concepts. Inevitably the types of precise measures needed for such algorithms include distortions that often reflect structural inequalities and related factors. These issues may be endemic among many industry algorithms across many industries.

4.3.4.2 Use in Court

Depending on the country, AI software is being used in courts. One relatively simple example is the use to determine the order in which cases are brought up to a judge, making use of information on the severity of cases, prior convictions, and more, in order to make a court's work more efficient (Lin et al. 2019). A more impactful use is in supporting judges to determine whether an inmate gets released on probation or not. A 2016 study by ProPublica found that the COMPAS system exhibited a systematic bias against African-American defendants in Broward County, Florida in terms of assess recidivism risk (Angwin et al. 2016).

This case generated considerable controversy. The developers of COMPAS (Northpointe, now Equivant) in their response to ProPublica's statistical analysis argued that ProPublica had "focused on classification statistics that did not take into account the different base rates of recidivism for blacks and whites" (Dieterich et al. 2016). This is a highly technical argument. Several commentators have observed there are multiple conceptions of "fairness" in the machine learning literature. With particular reference to the Broward County recidivism rates it has been demonstrated mathematically that "an instrument that satisfies predictive parity cannot have equal false positive and negative rates across groups when the recidivism prevalence differs across those groups" (Chouldechova 2017).

In many of these cases, countermeasures are not easily devised. There are multiple conceptions of fairness which might conflict. One might be faced with a possible trade-off between fairness, on the one hand, and accuracy on the other. In fact, some argue that efforts to 'blind' algorithms to objectionable categorical information, such as race, may be harmful and a better approach would be to alter how we use machine learning and AI (Kleinberg et al. 2018). While one certainly wants to avoid systematic biases in their AI programming, data should also reflect the actual situation, i.e. "what is the case", otherwise these data will not be very useful for drawing further conclusions and acting on their basis.

It is critical to appreciate the limits of classifications based on statistical models. While much of the Broward County recidivism controversy centred on the difference between false positives between blacks and whites, one study found the underlying predictive accuracy of COMPAS was only around 65%. The authors observed:"these predictions are not as accurate as we might want, particularly from the point of view of a defendant whose future lies in the balance" (Dressel and Farid 2018).

4.3.5 Explicability

The principle of Explicability states that it shall be possible to explain why an AI system arrived at a certain conclusion or result.

Explicability is not to be equated with transparency. While some call for maximal transparency of programs and codes, when it comes to AI systems this might not solve a problem and might even create new problems. Suppose software containing millions of lines of code is made transparent, what would be the benefit of this? First, the software would probably not be intelligible to non-experts, and even experts would struggle with what it means. Second, maximal transparency of software might create a risk vis-a-vis competitors, and hinder further investment in this sector. Due to considerations like these, some parts of the discussion have switched to the concept of “explicability”.

Explicability, as Floridi et al. (2018) define it, means both intelligibility and accountability. It is desirable in moral applications that those using AI systems or whose interests are affected by AI systems can “make sense“ of the precise way in which an AI made a particular decision. Intelligibility means that the workings of the AI can be understood by a human. The system is not a mysterious “black box” whose internals are unintelligible. Someone, even if only an experienced programmer, can understand the system’s workings and explain it to judges, juries and users.

The European Union implemented a legal requirement for the “right to information” (originally called the “right to explanation”) within the framework of the General Data Protection Regulation. This holds that people whose interests have been affected by an algorithmic decision have to right to have the algorithm explained the decision to them. Obviously, this poses a challenge for some “inscrutable” machine learning methods such as neural networks. There are some who are untroubled by “inscrutability” in machine learning. They take the view that they can test the system empirically. So long as it works in practice, they do not care if they cannot understand or explain how it works in practice (Weinberger 2018).

For many machine learning applications this may be fine. Yet, in morally charged situations that might be subject to judicial review, there will be a requirement to explain the decision. There may also be a requirement to justify the decision. A key element of moral functioning is not just doing the right thing but justifying what the agent did is right. Moral justification cannot be based on an “inscrutable” black box. However, there is ongoing research into “explainable AI” which seeks to generate explanations for why neural networks make the decisions they make (Wachter et al. 2017). It may be that such research eventually enables machine learning to generate adequate explanations for decisions it makes.

Even so, in practical terms, the use of the COMPAS system to assess recidivism risk and thus affect prospects for probation has been tested in court. In the case of Loomis versus Wisconsin, the plaintiff argued that he was denied “due process” because of the proprietary nature of the COMPAS algorithm meant that his defence could not challenge the scientific basis on which his score was calculated. However, his appeals failed. The judges held that sentencing decisions were not entirely based on COMPAS risk scores. Judges could not base sentences on risk scores alone but could consider such scores along with other factors in making their assessment of recidivism risk.

Accountability at its most basic can be provided in the form of log files. For example, commercial airlines are required to have a flight recorder that can survive a crash.

In the event of a plane crash, the flight recorders can be recovered. They facilitate the investigation of the crash by the authorities. Flight recorders are sometimes called “black boxes”. The log files generated by the system that can explain why it did what it did. This data is often used to retrace the steps the system took to a root cause and, once this root cause if found, assign blame.

4.4 Conclusion

To sum up, given a free choice, users will accept systems if they trust them, find them useful and can afford them. Businesses therefore have an incentive to make systems people trust. Trust involves a very complex cluster of concepts. To trust a system users will need to be confident that the system will do good to them or for them. That is it will display beneficence towards them. They will need to be confident the system will not do bad things to them such as harm them or steal from them or damage their interests in some other way (e.g. breach their privacy or cause them embarrassment).

They need to be confident the AI will not compromise their autonomy. This is not to say robots and AIs will never say no to humans. It is merely to say they will only say “no” if there is a good reason. For example, if the human wants to do something wrong with the robot, a morally intelligent robot could (in theory) refuse to obey the human.

People will need to be confident that the system can act justly within its functional scope. If there are clear laws and regulations and no moral controversies this is much easier to implement. Currently, due to the lack of agreement on moral theory, such functional scopes are narrow. Even so, there are numerous practical applications that are morally charged and yet adequately handled by AIs.

Finally, to trust machines, they will need to be explicable. For reasons of intelligibility and accountability, AIs will need to keep logs and explanations as to why they do what they do.

Discussion Questions:

- What risks would you take with an AI and what risks would you not take? Explain.
- Try to think of ways in which explicability of algorithms or AI could be implemented or operationalised. How might an AI explain to different people?
- We expect a robot or AI system to be more fair and unbiased than a human being. Are there limitations to this expectation? Discuss.

Further Reading:

- Wendell Wallach and Colin Allen. *Moral machines*: Teaching robots right from wrong. Oxford University Press, 2008. http://www.worldcat.org/oclc/1158448911.
- Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds and Machines* 28(4):689–707, Dec 2018. ISSN 1572-8641. Doi: 10.1007/s11023-018-9482-5. URL https://doi.org/10.1007/s11023-018-9482-5.

Chapter 5
Responsibility and Liability in the Case of AI Systems

This chapter discusses the question of who is responsible in the case of an accident involving a robot or an AI system that results in some form of damage. Assigning liability is challenging because of the complexity of the situation and because of the complexity of the system. We discuss examples of an autonomous vehicle accident and mistargeting by an autonomous weapon in detail to highlight the relationship of the actors and technologies involved.

People are not concerned with liability when everything is going fine and no one suffers harm. This concept comes into play only when something goes wrong. Some damage is done, people get hurt, property gets stolen. Then people want to know who is responsible, who gets blamed and who is liable for damages or compensation.

What holds for the non-digital world also holds for AI. Only when these technologies cause problems will there be a case for holding some person or entity responsible or liable. And these problems might be serious problems, i.e., problems that are encountered not just in a single case, but in many instances, in a systematic way. So what does "go wrong" mean in the case of AI technologies? Let's discuss two examples: the crash of an autonomous vehicle and mistargeting by an autonomous weapon.

5.1 Example 1: Crash of an Autonomous Vehicle

Imagine a vehicle driving in autonomous mode. Let's make this more specific: What does "autonomous" mean in this case? The European guidelines (and similarly the US NHTSA guidelines) distinguish between 6 levels of automated driving, from simple driver assistance systems which allow the car to brake or accelerate automatically to

C. Bartneck et al., *An Introduction to Ethics in Robotics and AI*,
SpringerBriefs in Ethics, https://doi.org/10.1007/978-3-030-51110-4_5

fully autonomous vehicles which do not require a driver at all (see Sect. 10.1). The crucial leap happens from level 2 to level 3. On level 3 a driver is not required to monitor the functions of their car at any time. This implies that a car has to be able to deal with a broad range of cases, including critical situations, on its own. The vehicle needs to do this at least for a certain amount of time. Car manufacturers often set this time frame to 10 s before requiring the driver to take control again.

So if an accident happens during the time that the car was in control, who would be held responsible or liable? The driver was not in control, not even required to do so. The technology did just as it was programmed to do. The company maybe? But what if they did everything possible to prevent an accident—and still it happened?

5.2 Example 2: Mistargeting by an Autonomous Weapon

In Chap. 11 we talk about the military use of artificial intelligence in more detail. Here, we will focus our attention on only the liability aspects of autonomous weapons. In the military context the system in question is typically an autonomous weapon. International Humanitarian Law (IHL) permits combatants to kill, wound and capture enemy combatants. They are also allowed to destroy enemy war materiel, facilities and installations that support the enemy war effort. However, combatants are not allowed to kill non-combatants unless they are actively engaged in the enemy war effort. Thus it is lawful to bomb an enemy factory, phone exchange or power station. Indeed, it is lawful to kill civilians assisting the enemy war effort if they are working in factories making military equipment that contributes to the war effort. However there are many objects that are specifically protected. One cannot bomb a hospital or school or attack an ambulance. In this example, we discuss the questions of blame, responsibility and liability with respect to an autonomous weapon killing a person it is not permitted to kill under IHL.

Let us suppose there is a war between a Red state and a Blue state. The commander of the Red Air Force is Samantha Jones. It so happens, she has an identical twin sister, Jennifer Jones, who is a hairdresser. She makes no direct contribution to the war effort. She does not work in an arms factory. She does not bear arms. One day during the war, Jennifer takes her child to a kindergarten as she does every day. Blue has facial data of Samantha but knows nothing about her twin sister, Jennifer. An "autonomous" drone using facial recognition software spots Jennifer taking her daughter to kindergarten. It misidentifies Jennifer as a "high value target" namely her sister, Samantha. Just after she drops her child off at the school, the drone fires a missile at her car, killing her instantly.

In this situation, who or what is to blame for the wrongful killing of Jennifer? Who or what should be held responsible for this wrong. Who or what is liable to be prosecuted for a war crime for this killing? Who or what would be punished?

First let's clarify what would happen if the drone was remotely piloted by human pilots. In the human scenario, the pilots also use facial recognition software. They manoeuvred the drone into good position to get clear images. They ran several images of Jennifer through their facial recognition system. They also looked at the photos of Samantha Jones they had themselves. All the images identified the image as General Samantha Jones. None identified the image as Jennifer Jones. This is because the facial recognition system only had images of high values targets, not of the entire enemy population. They were following orders from General Smith of the Blue Air Force to seek and destroy his counterpart in the Red Air Force.

5.2.1 Attribution of Responsibility and Liability

To blame an agent for a wrong, we must first establish the causal chain of events that led to the wrong. Also, we must understand who or what made the decisions that resulted in the wrong. In the human case, the humans relied on software to identify the target. They also used their own senses. The humans pressed the trigger that fired the missile. They did so because the software identified the target. The mistake was made because the software had incomplete information. There was an intelligence failure.

Is there much difference (if any) with regard to the Autonomous Weapon System (AWS) case? The AWS was following the same rules as the humans did. It fired for the same reason as the humans did. A computer database of facial images returned a match. Jennifer Jones did look like her sister and that is what led to her death.

There are one or two differences but both cases are very similar. Does it make sense to “blame” the AWS? Does it make sense to “blame” the humans? The fault lies in what was **not** in the facial recognition system. If Jennifer Jones was in the system the system might have reported that this image could be either Jennifer or Samantha. Then the humans would have to make a call as to what to do. With the facial recognition reporting “that is Samantha” and no one else, then a human would most likely assume they had their target too.

5.2.2 Moral Responsibility Versus Liability

We might well ask who is morally responsible for the wrongful death of Jennifer? We might consider the following people: the political leader who declared war, the pilot who fired the missile in the human case, the programmers who programmed the drone to fire missiles in the AWS case, or the intelligence officers who missed Samantha’s sister. In reality, these are all “contributing factors” to the death of Jennifer. However, we can make some distinctions between moral responsibility and legal responsibility (i.e., liability).

Legal responsibility has to be proven in a court. If the human pilots were to be subject to a court martial for the unlawful killing of a civilian, then the facts would have to be proven to a standard beyond reasonable doubt. Certainly, the defence would argue the human pilots had a “reasonable expectation” that the facial recognition system was accurate. Blame might be shifted to the intelligence officers who failed to find and load images of Jennifer into the system, or perhaps to the vendors of the facial recognition system. On these facts it is hard to see the pilots being found guilty of a war crime. Their intention was clearly to kill an enemy general. This would be a legal act in the context of war.

It has been said that law floats on sea of ethics. It is easy to imagine that General Jones would be furious to learn of the death of her sister. In the human case, she would naturally hold the pilots who fired the missile responsible. Perhaps she would hold the officer who ordered the mission responsible. Perhaps she would hold the political leadership of the Red State responsible. Regardless of who General Jones thinks is responsible, to assign legal responsibility, evidence would need to be presented in a court or legal proceeding. By bringing such proceedings, those who commit wrongs can be held accountable and those who do wrong can then be punished.

This is fine for humans but in the case of an AWS what does it mean to "punish" a system that cannot feel or suffer? If the AWS only does what it is programmed to do or acts on the basis of the data humans input into it, how can it be "responsible" for killing the wrong person? A pragmatic remedy is simply to assign responsibility for the actions of an AWS to a human or the state that operates it. There is an established legal concept called "strict liability" (described in the following section) which can be used in this regard.

5.3 Strict Liability

Product liability is the legal responsibility to compensate others for damage or injuries that a certain product has caused. Product liability is a concept that companies around the globe are familiar with and have to take into account. Clearly, when a company or one of its employees causes harm to others by neglecting their duties, the company will be held liable.

In some cases, there may not even have been a concrete wrongdoing. In the case of "strict liability," a company or a person can also be held liable even if they did not do anything wrong in the strict sense. For example, if someone owns a cat, and this cat causes damage to someone else's property, the owner is held liable in this sense. Or, a technique that has many beneficial consequences might also have some negative ones. For example, while vaccination is in general beneficial for the members of a society, there might be some cases where children suffer bad consequences from vaccines. These must be compensated for within the framework of strict liability. In the US the National Vaccine Injury Compensation Program provides this compensation.

A similar situation might arise in the context of autonomous vehicles or AI technologies in general, since a concrete (or even an abstract) harm may have been neither intended nor planned (otherwise it would have been a case of deliberate injury or fraud). Still, strict liability would be the framework within which to deal with these issues. Moreover, in many cases of AI technologies, harm might not have been foreseeable. This might even be a signature characteristic of these technologies, since they often operate in ways which are, in a sense, opaque to observers. Even to the programmers themselves it is often not clear how exactly the system arrived at this or that conclusion or result.

In the case of autonomous vehicles a several shifts have been made to move the liability from the driver or car owner to the car manufacturer or, if applicable, the company operating or developing the software. This shift was made possible through the adaptation of the Vienna Convention on Road Traffic. More specifically through its update from the 1968 version to the 2014 that took effect inf 2016 (United Nations 1968).

In the German Ethics Code for Automated and Connected Driving (Luetge 2017), this step is explicitly being taken, for those cases where the car's system was in control. This implies, among others, that monitoring (black box) devices will have to be installed in those cars which clearly record who was in control at each moment the driver or the car.

5.4 Complex Liability: The Problem of Many Hands

In the AWS case we can see that "many hands" may be involved in a wrongful death. General Blue ordered the mission which went wrong. The pilots confirmed the target based on faulty intelligence and pressed the button that launched the missile. The facial recognition system made a decision based on incomplete data. Those who designed the system will say the fault lies with those who input the data. How will a judge determine who is responsible?

In a complex liability matter it is often the case that no one is found to be at fault. Intuitively, people want to blame someone but in complex events, often there is no single person who can be blamed. In such cases, the normal legal solution is to assign blame to a collective entity. In this case, blame would be assigned to the Blue State not any particular individual. Strict liability can be assigned to States operating an AWS. Even if no person deserves blame, the collective entity is held responsible and sanctioned accordingly.

5.5 Consequences of Liability: Sanctions

What sanctions would eventually be imposed must be left open at this point. In many cases, there might be a focus on sanctions against a company rather than against individuals, since programmers will have to work with a certain company strategy or policy. Corporate liability will be a key solution here, as in the case of fines against Volkswagen (around EUR 25 billion in 2016) or Bank of America (around USD 16 billion in 2014). This shows that the sanctions can actually be very substantial.

Discussion Questions:

- Describe the differences between moral and legal responsibility in your own words.
- If an AI does the wrong thing, who would you blame? The software? The programmer(s)? The company making it? Explain.
- How would you compensate a person that is severely injured by an autonomous car? Give reasons.

Further Reading:

- Robin Antony Duff. Answering for crime: *Responsibility and liability in the criminal law*. Hart Publishing, 2007. ISBN 978-1849460330. URL http://www.worldcat.org/oclc/1073389374
- Martha Klein. Responsibility. In Ted Honderich, editor, *The Oxford companion to philosophy*. OUP Oxford, 2005. ISBN 978-0199264797. URL http://www.worldcat.org/oclc/180031201
- Christoph Luetge. Responsibilities of online service providers from a business ethics point of view. In Mariarosaria Taddeo and Luciano Floridi, editors, *The Responsibilities of Online Service Providers*, pages 119–133. Springer, 2017. Doi: 10.1007/978-3-319-47852-4 7. URL https://doi.org/10.1007/978-3-319-47852-47.

Chapter 6
Risks in the Business of AI

This chapter discusses the general risks that businesses face before considering specific ethical risks that companies developing AI systems and robots need to consider. Guidance on how to manage these risks is provided. It is argued that companies should do what is ethically desirable not just the minimum that is legally necessary. Evidence is given that this can be more profitable in the long run.

AI is not just a technology that concerns engineers and technical staff. It is a matter of business. Risks associated with advanced technology are likely to increase enormously in the years to come. Doing business with AI comes with many risks. Among the most significant are the risks associated with severe accidents. Take the example of autonomous vehicles, an AV out of control might cause great harm to other people and property. An AI system for controlling the power grid or other infrastructure functions might cause enormous damage if control is lost.

Such risks are not a matter for individuals only. They have consequences for companies, which can be critical and put them out of business entirely. Therefore, many large companies have entire departments for risk management. It is not our purpose here to discuss the more classic risks. Such risks include country risks, political risks, currency risks and related risks. These are well-known and well-understood by companies and insurers. However, the largest multinational global companies, who have the greatest number of skilled personnel and advanced scientific capabilities, are becoming increasingly aware of ethical risks. The more obvious ethical risks are corruption, discrimination or systematic abuse of human rights. The development of advanced technology capable of psychologically influencing people is also a risk. In the globalised world however, such risks might turn into economic risks eventually. There are several mechanisms through which this can occur:

C. Bartneck et al., *An Introduction to Ethics in Robotics and AI*,
SpringerBriefs in Ethics, https://doi.org/10.1007/978-3-030-51110-4_6

1. **Ethical risk to reputation**
   The reputation of a company may get damaged significantly. This is something not to be underestimated. The brand of large companies is often their most important asset. For example, in 2018, the value of the brand Apple was estimated to be USD 182 billion by Forbes. BMW's brand was valued at USD 31.4 billion. It would be foolish for such a company to damage their brand for very short-run purposes.
2. **Ethical risk to stock price**
   The stock price of a company might be affected greatly by ethical scandals. BP, for example, suffered great value to its stock price as a result of the Deepwater Horizon disaster in the Gulf of Mexico. In the month following this crisis, its stock price fell from USD 60 to USD 27 in a month. A fall in stock price can be something that destroys careers or leads to takeovers of the entire company. So this is something successful CEOs will seek to avoid.
3. **Ethical risk of legal fines**
   Finally, underestimating ethical risks can lead to significant legal fines. This is something that a number of large companies had to learn the hard way. For example, in the aftermath of the Deepwater Horizon scandal, BP has been forced to pay more than USD 65 billion in damages, fines and clean up costs. Also, Siemens had, before 2006, completely underestimated the issue of systematic corruption within their company. This risk was not regarded as significant, as it did not result in economic problems. However, eventually Siemens had to pay around EUR 2 billion in legal fines, which is a sum that even such a large company would prefer to avoid. As of 2017, according to CNN, the cumulative costs of the Dieselgate scandal for VW are in the order of USD 30 billion.

6.1 General Business Risks

Some of the key risks faced by companies are:

6.1.1 Functional Risk

Functional risk is simply the risk of the functionality of the system failing. For example, in any software and hardware system there is always the risk that the system may fail when released to the general public. A component may fail in certain unanticipated situations. Sometimes an update patch for a software system may fail and cause the "breaking" of the system itself or another system it interacts with it. Updates often break integration between systems provided by different vendors.

6.1.2 Systemic Risk

Systemic risks are risks that affect a whole system. Systemic risk in finance, for example, is the risk of the entire financial system collapsing. The Global Financial Crisis of 2007–08 was caused by widespread loan defaults in the US subprime market. Subprime loans are given to people with weak credit ratings. Defaults stressed the major US mortgage lenders Fannie Mae and Freddie Mac and many home owners abandoned their houses. This led to a collapse in confidence that exposed inadequate risk models. One of the causes of the collapse of Lehman Brothers was wholesale underpricing of the risk of mortgage defaults and lowering real estate prices. This was due to faulty financial models. When Lehman Brothers filed for bankruptcy in 2008, banks stopped trusting each other and started to refuse to lend to each other. The global financial system required massive bailouts from governments.

Sophisticated AI systems are widely used in finance. Indeed automated trading systems are widely used by investment bankers and financial traders. The Flash Crash of 2010 which lasted for 36 min caused the US stock market to lose 9% of its value. High-frequency traders that used AI systems were involved in this incident. There are fears that use of such systems might "blow up" the market and cause a recession or even a depression.

6.1.3 Risk of Fraud

Computer systems have been used to perpetrate fraud. One of the largest was the Dieselgate scandal. Volkswagen deliberately designed their emissions reduction system to only function during laboratory tests. As a result of this criminal deception their cars passed tests in labs but emitted up to forty times these volumes on the road. Even so, Volkswagen promoted their fraudulently obtained "green" credentials. The software used to detect the laboratory test was relatively simple and used parameters, such as the steering wheel inclination (Contag et al. 2017). Still, this is an example of how a computer system can be used in large scale frauds.

6.1.4 Safety Risk

Robots that control industrial production present physical risks to the people that work around them. People may be harmed by force of collisions or harmed by the objects that a robot carries or moves. The Deepwater Horizon explosion is an example of a situation where a system failure led to a catastrophe that killed eleven workers. Similarly, autonomous vehicles have, on occasion, caused accidents that have led to fatalities. Uber and Tesla AVs have been involved in fatal collisions.

6.2 Ethical Risks of AI

There are ethical risks for AI and robotic technologies. Managing them becomes a crucial task for managers in the globalised economy. With this in mind what are the risks of AI technologies and robotics? Some of the main ethical risks are.

6.2.1 Reputational Risk

Systems that appear biased or prejudiced can cause great reputational damage. Hot topics in this area include facial recognition and loan approval systems. However a spirited campaign of "testing, naming and shaming" has dramatically increased the ability of commercial systems to correctly recognise the faces of females and minorities (Buolamwini and Raji 2019).

One of the first major scandals in the still young social media economy is the Cambridge Analytica scandal, which greatly damaged the reputation of Facebook. Facebook supplied data obtained by Cambridge Analytica and which was used for targeted political advertising.

6.2.2 Legal Risk

Legal risk encompasses situations where a system becomes too successful and is viewed as causing an anti-competitive environment. This situation may attract the attention of governmental regulators capable of levying large fines. In 2004, Microsoft, for example, was fined EUR 497 million by the European Union for anti-competitive behaviour. In 2018, the European Commission imposed a record fine of USD 5 billion on Google for antitrust violations of the Android technology.

6.2.3 Environmental Risk

Similarly system failures can cause environmental disasters. The Bhopal disaster in India, for example, was the result of a gas leak at an Union Carbide factory in Bhopal India. The leak resulted in an explosion and release of methyl isocyanate. The immediate explosion caused the death of nearly 4,000 people. The gas release however may have harmed more than half a million people. Several thousand would eventually be determined to have suffered permanent damage from exposure to the toxic substance. Union Carbide eventually paid approximately USD 470 million in fines and restitution.

6.2.4 Social Risk

Social risk includes actions that may include the people, society or communities around the business. These risks may include increased traffic, noise pollution, and issues related to worker morale. Social risks related to AI may include technology induced increased social isolation, increased inequality, and local community issues surrounding the acceptable uses of technology and AI. For example, the use of Google goggles generated numerous issues surrounding the privacy and the use of an integrated camera in private places.

The examples above briefly introduce some of the more ethically charged risks that companies must manage to survive in the modern world. There are numerous other risks besides these (currency risk, commercial risk etc.) which we do not cover here.

6.3 Managing Risk of AI

One important question is how these risks can be assessed, and even quantified, if possible. Usually, it makes a big difference especially to large companies, if a risk can be estimated in money. We have already mentioned the loss in the value of a company brand, in stock price and in legal fines, but there are also other attempts in quantifying reputational risk, by a mix of asking external experts, doing surveys with employees or other measurements. Again, this leads to a value which will then have consequences for a company, internally and externally.

For AI companies as well as in others, there are several ways of dealing with these risks. First there are legal methods of establishing or improving existing safety regulations. For example, in the case of autonomous cars, laws are being adjusted to accommodate the inclusion of these technologies. At the same time, industry best practices are also being developed and implemented.

Second, regulations surround ancillary industries may need to be developed. For autonomous cars, this includes changes to insurance regulations, perhaps including the introduction or updating of compulsory insurance to meet the needs of this rapidly developing technological.

Finally, many companies will certainly, voluntarily or not, be compelled to go beyond the minimum legal standards for reasons of their own ethical values and addressing risks. Being a good corporate citizen implies controlling risks—in your own company as well as in the supply chain. For example, when the Rana Plaza building in Bangladesh collapsed in 2013, the public and many consumers viewed this as a problem of large multinational companies in the clothing industry and not just as a local problem. Similar viewpoints have been held in the case of Apple or Samsung as being responsible for child labour in the cobalt mines used in their supply chain.

DuckDuckGo, a search engine competing with Google, heavily advertises the fact that is does not collect data from people. As the saying goes in AI, “if you are not paying, you’re the product.” Personal data has considerable value to those seeking to target their marketing. DuckDuckGo is seeking to differentiate themselves from Google in the search engine market by not collecting user data. Whether it can displace Google from its dominant position in search and targeted advertising marketing by doing this remains to be seen. The value placed on personal data for the purposes of AI has been discussed in Sect. 8.2.

6.4 Business Ethics for AI Companies

Business ethics in a global environment has, to a significant part, become an instance of risk management for AI companies. Mechanisms like reputation, stock prices and legal fines exhibit a business case for ethics and may help to convince companies to be active in this area.

Let’s consider another example: The Malampaya project was a pipeline project by the Shell corporation in the Philippines in 1998 (Pies 2010). There were three possible ways to build this pipeline. First, it could have been built primarily on land, which would have been the cheapest way, but one which would have put the biodiversity of the region at great risk.

The second option would have been to build the pipeline in a direct, more expensive underwater route, but that route would have gone through a territory that was considered holy by indigenous people (who, however, had no legal claim to it).

Still, Shell eventually decided to go for a third, and much more costly option: to build the pipeline underwater, but on a different and much longer route around an island, which led it far away from the holy territory. There was no legal obligation for Shell to do so, and the additional cost was around 30 million US dollars. Although, this amount is not a significant cost for a company like Shell, it still represents a cost that would have been avoided if possible.

The deciding factor at the time probably was the aftermath of the 1995 Brent Spar Case: Shell tried to sink an unused oil platform in the North Sea, resulting in protests from Greenpeace and others and even in an initial boycott by consumers. Eventually, Shell abandoned their plan (though Greenpeace was later forced to admit that there was much less residual oil in the platform than it had alleged). Shell may have been trying to avoid a similar scandal.

What is interesting, however, is that in 2007 independent calculations by the World Resources Institute came to the conclusion that what seemed at first to be the most expensive option actually turned out to have been the more economical one, when taking into account other costs such as the costs of delayed completion. The indigenous people in question, while not having been able to stop the project entirely, still could have significantly delayed it by going to court. The delays would have resulted in penalties and other costs. Therefore, the overall calculation of all these resulted in the chosen option actually being proven to have made economic sense.

Not all cases have happy endings. Still, it is valuable for companies to consider examples such as this. These case studies may encourage companies to include ethics and environmental information in their calculates and to invest money in Corporate Social Responsibility initiatives. The same might hold in the area of AI, these companies should improve their inclusion of ethics related considerations, calculating over the long run, and consider systematically the interests of others who may be relevant for their future sustainable operations.

From a more general point of view, taking on ethical responsibilities (without waiting for changes in the legal framework) has become both a necessity as well as a competitive advantage for companies. Studies show that it helps raise profits (Wang et al. 2016), secure the capability for innovation, improve market position (Martinez-Conesa et al. 2017), improve risk management (Shiu and Yang 2017) and customer satisfaction (Lins et al. 2017) and also motivate employees and recruit new talent (Saeidi et al. 2015).

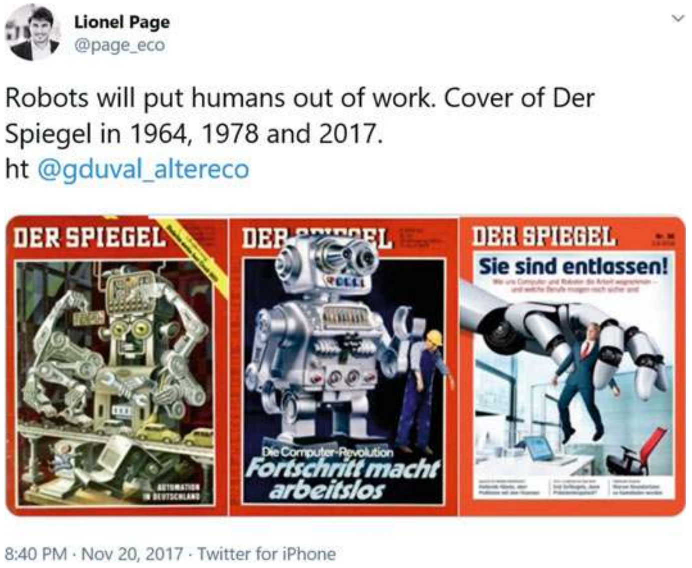
One word of caution seems appropriate: Companies can do a lot of things on their own, but they cannot do everything. In some sectors of the industry, it will be more difficult for a company to take responsibility on their own. Consider the banking sector. Broadly speaking, there is not as much room for direct action by banks as in other industries, simply because the banking sector is very tightly regulated.

The situation in AI could become similar. As a product developed by AI may be capable of making important decisions in the world and as it is difficult or impossible to argue or appeal to these programs or algorithms, we will need specific ethical guidelines for AI. This does not necessarily mean detailed regulation. But rather guidelines that are the subject and focus of societal discussion. Hence, they need to address important problems that people care about—even if they might not be able to solve each of these problems.

6.5 Risks of AI to Workers

Many writers have predicted the emergence of “mass technological unemployment” resulting from the increasing use of AI and robots. The pessimistic line is that humans will all eventually be made redundant by machines. Numerous reports have made alarming predictions that huge numbers of human jobs will be automated away (Frey and Osborne 2017). For example, there are around three million truck drivers in the United States. If a viable autonomous truck were available tomorrow, in theory, all three million truck drivers could be made redundant very quickly. There are many stories about how AI and robotic will make everyone redundant. However, these stories are based on uncritically accepted assumptions.

For example, Frey and Osborne’s projections have been widely criticised. They predicted that nearly half the jobs in the workforce were vulnerable to automation over the next 20–30 years. However, an OECD report using different assumptions arrived at a much less alarming figure of 9% (Arntz et al. 2016).



Lionel Page
@page_eco

Robots will put humans out of work. Cover of Der Spiegel in 1964, 1978 and 2017.
ht @gduval_altereco

8:40 PM · Nov 20, 2017 · Twitter for iPhone

Fig. 6.1 Der Spiegel covers on mass unemployment

Certainly, if the predictions of looming "mass technological unemployment" were true, we would expect to see lengthening unemployment lines in the most advanced technological nations. However, prior to the COVID-19 pandemic, employment in the United States is currently at 30 year lows. So, it would seem that new jobs are being created to replace those lost to automation.

Another myth-busting consideration is that headlines predicting the replacement of workers by robots are not new. The German current affairs magazine *Der Spiegel* has run three cover stories predicting that robots would put people out of work, in 2017, 1978 and 1964 as shown in Fig. 6.1.

The main counter-argument to the claim that AIs and robots will put us all out of work is the fact that automation has been going on since the Industrial Revolution. Technological advancement does disrupt employment. This is undeniable. However, new technology means workers can do more with less and they can do more interesting and valuable things as wealth in a society increases. Historically, new jobs have been created to replace the old ones. If the predictions of mass technological unemployment were true, then we would expect to see high unemployment in technologically advanced states. There is no sign of this happening yet.

Discussion Questions:

- Are you worried about AIs and robots putting you out of work or excited by the possibilities of new technology? Explain.
- Give some examples of how bias could be introduced into an AI.
- Do you think there should be a compulsory insurance scheme for the risks associated with AVs?

Further Reading:

- Eric Bonabeau. Understanding and managing complexity risk. *MIT Sloan Management Review*, 48(4):62, 2007. URL https://sloanreview.mit.edu/article/understanding-and-managing-complexity-risk/
- A Crane and D. Matten. *Business Ethics. Managing Corporate Citizenship and Sustainability in the age of Globalization*. Oxford University Press, 2007. ISBN 978-0199697311. URL http://www.worldcat.org/oclc/982687792
- Christoph Luetge, Eberhard Schnebel, and Nadine Westphal. Risk management and business ethics: Integrating the human factor. In Claudia Klüppelberg, Daniel Straub, and IsabellWelpe, editors, *Risk: A Multidisciplinary Introduction*, pages 37–61. Springer, 2014. Doi: 10.1007/978-3-319-04486-6. URL https://doi.org/10.1007/978-3-319-04486-6.

Chapter 7
Psychological Aspects of AI

In this chapter we discuss how people relate to robots and autonomous systems from a psychological point of view. Humans tend to anthropomorphise them and form unidirectional relationships. The trust in these relationships is the basis for persuasion and manipulation that can be used for good and evil.

In this chapter we discuss psychological factors that impact the ethical design and use of AIs and robots. It is critical to understand that humans will attribute desires and feelings to machines even if the machines have no ability whatsoever to feel anything. That is, people who are unfamiliar with the internal states of machines will assume machines have similar internal states of desires and feelings as themselves. This is called anthropomorphism. Various ethical risks are associated with anthropomorphism. Robots and AIs might be able to use "big data" to persuade and manipulate humans to do things they would rather not do. Due to unidirectional emotional bonding, humans might have misplaced feelings towards machines or trust them too much. In the worst-case scenarios, "weaponised" AI could be used to exploit humans.

7.1 Problems of Anthropomorphisation

Humans interact with robots and AI systems as if they are social actors. This effect has called as the "Media Equation" (Reeves and Nass 1996). People treat robots with politeness and apply social norms and values to their interaction partner (Broadbent 2017). Through repeated interaction, humans can form friendships and even intimate relationships with machines. This anthropomorphisation is arguably hard-wired into our minds and might have an evolutionary basis (Zlotowski et al. 2015). Even if the designers and engineers did not intend the robot to exhibit social signals, users might still perceive them. The human mind is wired to detect social signals and to interpret

C. Bartneck et al., *An Introduction to Ethics in Robotics and AI*,
SpringerBriefs in Ethics, https://doi.org/10.1007/978-3-030-51110-4_7

even the slightest behaviour as an indicator of some underlying motivation. This is true even of abstract animations. Humans can project "theory of mind" onto abstract shapes that have no minds at all (Heider and Simmel 1944). It is therefore the responsibility of the system's creators to carefully design the physical features and social interaction the robots will have, especially if they interact with vulnerable users, such as children, older adults and people with cognitive or physical impairments.

To accomplish such good social interaction skills, AI systems need to be able to sense and represent social norms, the cultural context and the values of the people (and other agents) with which they interact (Malle et al. 2017). A robot, for example, needs to be aware that it would be inappropriate to enter a room in which a human is changing his/her underwear. Being aware of these norms and values means that the agent needs to be able to sense relevant behaviour, process its meaning and express the appropriate signals. A robot entering the bedroom, for example, might decide to knock on the door prior to entering. It then needs to hear the response, even if only non-verbal utterance, and understand its meaning. Robots might not need to be perfectly honest. As Oscar Wilde observed "The truth is rarely pure and never simple." White lies and minor forms of dishonesty are common in human-human interaction (Feldman et al. 2002; DePaulo et al. 1996).

7.1.1 Misplaced Feelings Towards AI

Anthropomorphism may generate positive feelings towards social robots. These positive feelings can be confused with friendship. Humans have a natural tendency to assign human qualities to non-human objects. Friendships between a human and an autonomous robot can develop even when the interactions between the robot and the human are largely unidirectional with the human providing all of the emotion. A group of soldiers in Iraq, for example, held a funeral for their robot and created a medal for it (Kolb 2012). Carpenter provides an in-depth examination of human-robot interaction from the perspective of Explosive Ordinance Disposal (EOD) teams within the military (Carpenter 2016). Her work offers an glimpse of how naturally and easily people anthropomorphise robots they work with daily. Robinette et al. (2016) offered human subjects a guidance robot to assist them with quickly finding an exit during an emergency. They were told that if they did not reach the exit within the allotted 30 s then their character in the environment would perish. Those that interacted with a good guidance robot that quickly led them directly to an exit tended to name the robot and described its behaviour in heroic terms. Much research has shown that humans tend to quickly befriend robots that behave socially.

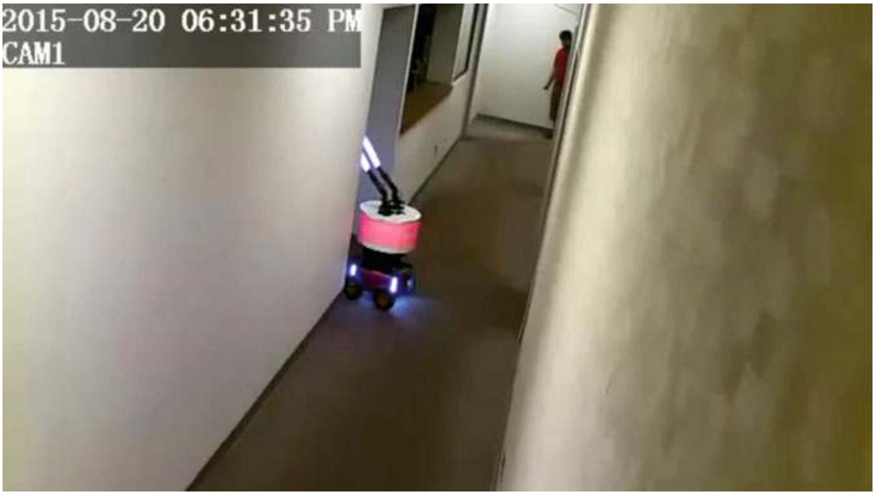


Fig. 7.1 Robot guiding people out of a building

7.1.2 *Misplaced Trust in AI*

Users may also trust the robot too much. Ever since the Eliza experiments of the 1960s, it has become apparent that computers and robots have a reputation of being honest. While they rarely make mistakes in their calculations, this does not mean that their decisions are smart or even meaningful. There are examples of drivers blindly following their navigation devices into even dangerous and illegal locations. Robinette et al. (2016) showed that participants followed an obviously incompetent robot in a fire evacuation scenario. It is therefore necessary for robots to be aware of the certainty of their own results and to communicate this to the users in a meaningful way (Fig. 7.1).

7.2 Persuasive AI

By socially interacting with humans for a longer period, relationships will form that can be the basis for considerable persuasive power. People are much more receptive to persuasion from friends and family compared to a car salesperson. The first experiments with robotic sales representatives showed that the robots do have sufficient persuasive power for the job (Ogawa et al. 2009). Other experiments have explored the use of robots in shopping malls (Shiomi et al. 2013; Watanabe et al. 2015). This persuasive power can be used for good or evil.

The concern is that an AI system may use, and potentially abuse, its powers. For example, it might use data, such as your Facebook profile, your driving record or your credit standing to convince a person to do something they would not normally do. The result might be that the person's autonomy is diminished or compromised when interacting with the robot. Imagine, for example, encountering the ultimate robotic car sales person who knows everything about you, can use virtually imperceptible micro expression to game you into making the purchase it prefers. The use of these "superpowers" for persuasion can limit a person's autonomy and could be ethically questionable.

Persuasion works best with friends. Friends influence us because they have intimate knowledge of our motivations, goals, and personality quirks. Moreover, psychologists have long known that when two people interact over a period of time they begin to exchange and take on each other subtle mannerisms and uses of language (Brandstetter et al. 2017). This is known as the Michelangelo phenomenon. Research has also shown that as relationships grow, each person's uncertainty about the other person reduces fostering trust. This trust is the key to a successful persuasion. Brandstetter and Bartneck (2017) showed that it only takes 10% of the members of a community to own a robot at which changes in the use of language in the whole community can take place.

More importantly, people might be unaware of the persuasive power of AI systems similar to how people were unaware of subliminal advertising in the 1950s. It is unclear who will be in control of this persuasive power. Will it be auctioned off for advertisers? Will the users be able to set their own goals, such as trying to break a bad habit? Unsophisticated people might be exploited and manipulated by large corporations with access to their psychological data. Public scrutiny and review of the operations of businesses with access to such data is essential.

7.3 Unidirectional Emotional Bonding with AI

The emotional connection between the robot or AI system and its user might be unidirectional. While humans might develop feelings of friendship and affection towards their silicon friends and these might even be able to display emotional expressions and emit signals of friendship, the agent might still be unable to experience any "authentic" phenomenological friendship or affection. The relationship is thereby unidirectional which may lead to even more loneliness (Scheutz 2014). Moreover, tireless and endlessly patient systems may accustom people to unrealistic human behaviour. In comparison, interacting with a real human being might become increasingly difficult or plain boring.

For example, already in the late 1990s, phone companies operated flirt lines. Men and women would be randomly matched on the phone and had the chance to flirt with each other. Unfortunately, more men called in than women and thus not all of the men could be matched with women. The phone companies thus hired women to fill the gap and they got paid by how long they could keep the men on the line. These professional talkers became highly trained in talking to men. Sadly, when a real woman called in, men would often not be interested in her because she lacked the conversational skill that the professional talkers had honed. While the phone company succeeded in making profit, the customers failed to achieve dates or actual relationships since the professional women would always for unforeseeable reasons be unavailable for meetings. This example illustrates the danger of AI systems that are designed to be our companion. Idealised interactions with these might become too much fun and thereby inhibit human-human interaction.

These problems could become even more intense when considering intimate relationships. An always available amorous sex robot that never tires might set unrealistic if not harmful and disrespectful expectations. It could even lead to undesirable cognitive development in adolescents, which in turn might cause problems. People might also make robotic copies of their ex-lovers and abuse them (Sparrow 2017).

Even if a robot appears to show interest, concern, and care in a person, these robots cannot truly have these emotions. Nevertheless, naive humans tend to believe that the robot does in fact have emotions as well, and a unidirectional relationship can develop. Humans tend to befriend robots even if they present only a limited veneer of social competence. Short et al. (2010) found that robots which cheated while playing the game rock, paper, scissors were viewed as more social and got more attributions of mental state compared to those that did not. People may even hold robots as morally accountable for mistakes. Experiments have shown that when a robot incorrectly assesses a person's performance in a game, preventing them from winning a prize, people hold the robot morally accountable (Kahn et al. 2012).

Perhaps surprisingly, even one's role while interacting with a robot can influence the bond that develops. Kim, Park, and Sundar asked study participants to either act as a caregiver to a robot or to receive care from a robot. Their results demonstrate that receiving care from a robot led participants to form a more positive view of the robot (Kim et al. 2013). Overall, the research clearly shows that humans tend to form bonds with robots even if their interactions with the robot are one-directional, with the person providing all of the emotion. The bond that the human then feels for the robot can influence the robot's ability to persuade the person.

Discussion Questions:

- Are there some things that you would rather discuss with an AI than a human? Create a list of general topics that might be easier to confess with an AI.
- Should robots always tell the truth, even if it results in socially awkward situations? List some situations that would be awkward.
- What is unidirectional emotional bonding? What makes it possible? Explain.

Further Reading:

- Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3-4):143–166, 2003. Doi: 10.1016/S0921-8890(02)00372-X. URL https://doi.org/10.1016/S0921-8890(02)00372-X
- Michael A Goodrich, Alan C Schultz, et al. Human-robot interaction: a survey. *Foundations and Trends in Human-Computer Interaction*, 1(3):203–275, 2008. Doi: 10.1561/1100000005. URL http://dx.doi.org/10.1561/1100000005.

Chapter 8
Privacy Issues of AI

This chapter sheds light on how private data is systematically collected, stored and analysed with the help of artificial intelligence. We discuss various forms of persistent surveillance at home and in public spaces. While massive data collection raises considerable ethical concerns, it is also the basis for better performance for AI systems.

8.1 What Is Privacy?

In its most basic form, privacy is the right to not be observed. People often act in a manner so as to increase their own privacy. For instance, shutting doors, wearing sun glasses, or clothing that is less revealing are simply subtle ways that humans actively moderate their own privacy.

Privacy is valuable for a number of important reasons: It allows people to make their own, non-coerced decisions, to better calculate their behaviour and be strategic in their social interactions, and also to take decisions and actions that do not conform to certain social norms.

Individual privacy has long had a contentious relationship with government. Just as individuals have typically fought for the right not to be observed, governments have often fought for the right to observe their own citizens.

Governments have long argued that "the need for privacy" argument may be used to cloak the planning and perpetration of crimes. Governments often argue that privacy must be curtailed in order to allow for proper law enforcement. Governments have argued that in some contexts, such as prisons, any expectation of privacy is not

C. Bartneck et al., *An Introduction to Ethics in Robotics and AI*,
SpringerBriefs in Ethics, https://doi.org/10.1007/978-3-030-51110-4_8

valid, but even in many public situations, most expectations of privacy need to be curtailed.

Governmental control over privacy has often been asserted to be a crime prevention method, but it can also serve the purpose of a general method of control. Limiting privacy provides governments with information about which individuals within a society need to be most controlled.

In the United States, questions about the right to privacy are fluid and subject to constant change as court decisions have, at times, awarded individuals more or less privacy. The United States Supreme Court has often considered privacy as a function of context, where a person's physical environment determines what privacy they should expect. But other considerations, such as the person's age or mental capacity may also be applied.

Companies and businesses are a third group that has traditionally been neutral in this debate. Companies try to perform a balancing act by weighing the needs and demands of customers against the needs and demands of government. Businesses often attempt to protect their own privacy and the privacy of their customers. Apple, for example, continues to deny government agencies systematic access to iPhones that have been locked, even if the data in the phone might help in the prevention or prosecution of crime (Holpuch 2016).

Companies may believe that if customers do not feel that their transactions with a company are private then they will no longer be customers. But companies must also rely on governments for a variety of policy, legal, and contractual obligations. Governments may use these powers as a method of persuasion to influence companies to hand over private data.

For centuries, this dynamic has existed, with groups not gaining much leverage over the other. Historically even the most authoritarian regimes had insufficient technological prowess to actually capture and keep track of private actions of their citizens. In the past 50 years this has changed. The development of computers and methods of mass data storage has allowed governments and business to collect, store, and process voluminous amounts of data, not only about its citizens but about all global citizens. Even more recently, the development of sensors capable of capturing data has broadened the horizon of where, when, how, and what data can be collected.

Finally, AI companies, long neutral in the debate over privacy, have begun to find value in the data that they have. They can now use AI methods to locate, model, and test potential products on populations of people. Using the psychological techniques discussed in Chap. 7, they can use powerful methods that exploit human tendencies for the purposes of improved advertising, marketing, and sales. The result is that the right to privacy is eroding. So is there any need for these data? Or is data collection simply done for its own sake?

8.2 Why AI Needs Data

Current robots and AI systems collect enormous quantities of data about their users. These data collections require considerable effort, and it might not be obvious at first sight what the benefit of this collection epidemic is.

Let’s consider the case of Facebook. Mark Zuckerberg was summoned to appear before a Senate Judiciary and Commerce Committee hearing on April 10, 2018. During the hearing Senator Orrin Hatch asked: “How do you sustain a business model in which users don’t pay for your service?”, to which Zuckerberg simply responded: “Senator, we run ads.”.

John Wanamaker famously pointed out that “Half the money I spend on advertising is wasted; the trouble is I don’t know which half.” What makes Facebook such an attractive platform for placing advertisement is that it knows a lot about its users and allows marketing experts to target specific groups. This targeting dramatically improves the efficiency and effectiveness of the ads. Less than the famous 50% of the advertisement gets wasted this way. The German online retailer Otto, for example, was able to reduce the cost per click by 38% through highly targeted advertisements (Kleinz 2018). Facebook users that previously ‘liked’ a picture of a friend in Hawaii were presented with advertisements for cheap flights to that destination.

While the advertising industry might be one of the greatest beneficiaries of private data, its use stretches far beyond this. A companion robot is likely to be perceived as more relatable and friendly if it knows and adapts to its specific user. A smart home might also learn when its inhabitants are present and adapt the heating to maximise comfort and energy efficiency. In summary, the more and better information about users is available, the better AI systems can learn from the data and adapt their behaviour.

8.3 Private Data Collection and Its Dangers

Gathering personal data has become dramatically easier with the arrival of certain key technologies, such as smartphones, surveillance cameras and of course, the Internet. These days it is in principle possible to track every step users take and every restaurant they visit. People take photos of the food they eat and post them online. Within the framework of the Self Optimisation movement people feverishly collect personal data in an attempt to change their lives for the better. Much of this data is now being uploaded to cloud computers, which has significantly increased the possibility for tracking private information.

Moreover, users of social networks voluntarily upload very private data and seem to deliberately ignore that by uploading data they often transfer the copyright of this data to the platform provider. Facebook and others *own* the data and use it and even sell it to others.

One of the factors of Google’s success in gathering personal data is that people cannot hide their interest when searching for information. While many would try to conceal delicate private issues they cannot search for information about this topic without entering the terms into the search box. Stephens-Davidowitz and Pinker (2017) analysed such personal search query patterns and identified a considerable group of Indian husbands that desire to be breast-fed. How Google will respond to

Fig. 8.1 Amazon Echo Plus uses Alexa (*Source* Amazon)

these results remains to be seen, but it does show that even our most intimate wishes and issues are being collected online.

Autonomous vehicles typically also report back telemetry data about the cars' performance, which in turn allows these companies to compile reports on a car's safety. Tesla, for example, compiles a quarterly report on the kilometres driven by its vehicles and whether the auto pilot was engaged.

8.3.1 Persistence Surveillance

Persistent surveillance is the constant observation of a person, place or thing. Within the military and policing fields, persistent surveillance is a commonplace technique for gathering information about an enemy or suspect. Yet, with the development of so-called digital assistants such as Amazon's Alexa (see Fig. 8.1) and Google Home, similar elements have become a product feature.

These systems stream audio data from the home to the parent company where the data is stored, collected, and analysed. Not only is this data quickly examined for requests from the product, but it might theoretically also be used to observe users and their environment in ways that are unknown to them. Sensitive to these concerns, Amazon has put a number of features in place to limit the data collecting ability of its devices. According to Amazon, features such as using the word "Alexa" to wake the device up prevents the device from being used as a means for persistent surveillance (Hildrenbrand 2018). Yet, as shown through the Wikileaks release of NSA documents, backdoors and vulnerabilities can be exploited in these types of technologies which could theoretically convert them into a means of persistent surveillance.

Another prominent example of private data collection is the "Hello Barbie" doll built by Mattel. It uses cloud-based speech recognition, conversation management and speech synthesis. While this is not unlike Alexa or Google Home, "Hello Barbie" was targeted at young girls. They were encouraged to talk with Barbie about their lives, and since they might not have been aware of the technical and social implications, they would likely have shared unfiltered thoughts with Mattel. Of course, Mattel vowed to respect the privacy of its customers, but it remains a question of trust whether parents believed them (Fig. 8.2).

Consumers, on the other hand, seem somewhat ambivalent towards privacy concerns about these devices. In a study conducted by Shields (Shields 2018), for instance, consumers report that they like the home security features of a digital assistant, yet also fear being spied on. Privacy-related events can cause citizens to catch on to privacy-related issues. A study by Anton et al. (2010) found that internet users had become more concerned about privacy even as their usage increased from 2002 to 2008.

In contrast to digital assistants which are willingly placed in the home by a person, persistent surveillance can also be conducted from a distance. Drone surveillance developed for the battlefield allows for continuous observation of individuals, i.e., the collection of information about individuals including the creation of social networks, and the creation of behavioural models that capture patterns of behaviour. In warfare, these social networks are used to distinguish non-combatants from enemy soldiers and to predict upcoming attacks. Recently, these practices have been adapted from the battlefield and applied to domestic locales. In 2005, Baltimore created a ground level surveillance system called CitiWatch which contained more than 700 cameras placed around the city.

In October 2014, CitiWatch expanded to create a database which privately owned surveillance cameras could contribute to on a voluntary basis. Later on, the Baltimore program began to include aerial surveillance from camera-laden air planes. On the one hand, Baltimore law enforcement officials claim that these programs work to reduce crime. On the other hand, privacy advocates claim that these programs greatly restrict privacy and inhibit personal freedom. Informal polls conducted by the Baltimore Business Journal and the Baltimore Sun found that 82% of people felt "comfortable" with the surveillance program "as long as it's keeping people safe." Moreover, an online Baltimore Sun poll found that 79% of people believed that the police department's level of secrecy around the program was acceptable (National Police Foundation 2017).

Similarly, airline authorities have considered continuously streaming the audio from airplane cockpits. Doing so would reduce the need for retrieving black box recorders. But pilots have hitherto resisted streaming audio from the cockpit on the

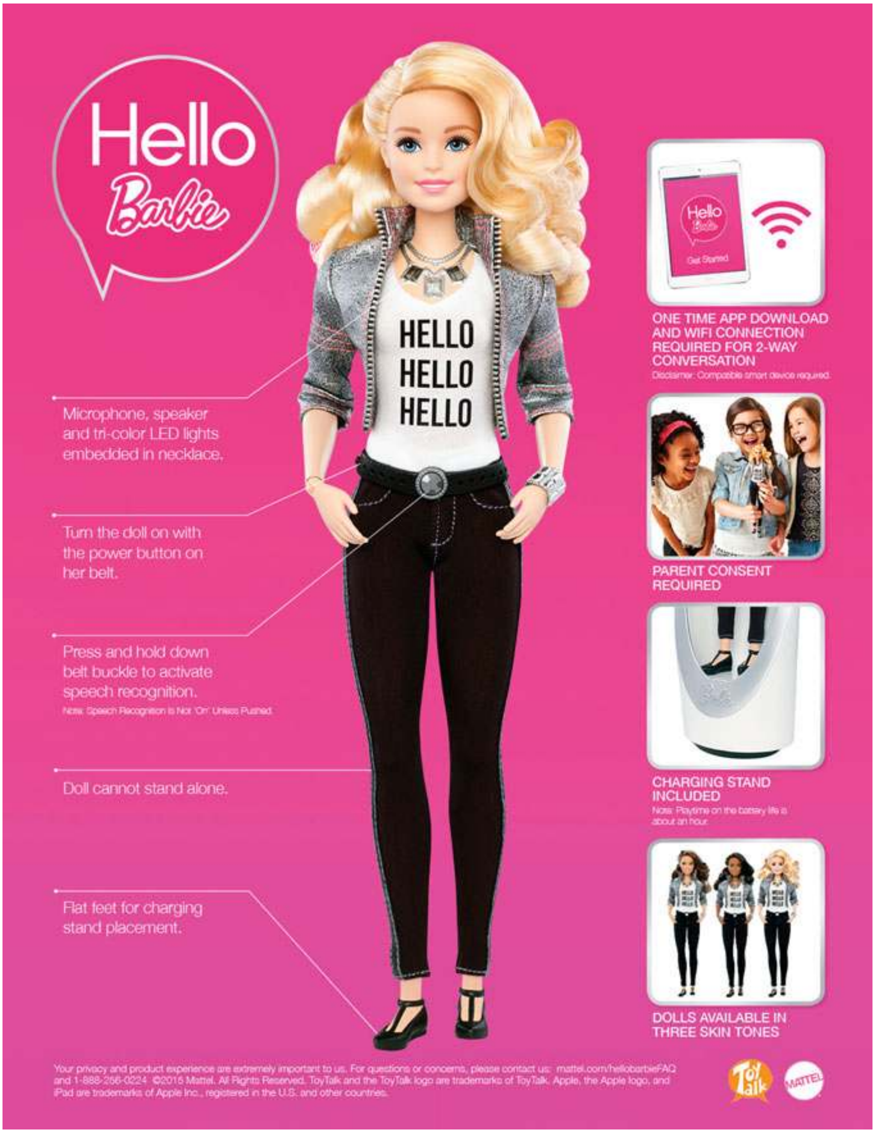

Hello
Barbie
HELLO
HELLO
HELLO
Microphone, speaker and tri-color LED lights embedded in necklace.
Turn the doll on with the power button on her belt.
Press and hold down belt buckle to activate speech recognition.
Note: Speech Recognition Is Not 'On' Unless Pushed.
Doll cannot stand alone.
Flat feet for charging stand placement.
Hello Barbie
Get Started
ONE TIME APP DOWNLOAD AND WIFI CONNECTION REQUIRED FOR 2-WAY CONVERSATION
Disclaimer: Compatible smart device required.
PARENT CONSENT REQUIRED
CHARGING STAND INCLUDED
Note: Playtime on the battery life is about an hour.
DOLLS AVAILABLE IN THREE SKIN TONES
Your privacy and product experience are extremely important to us. For questions or concerns, please contact us: mattel.com/hellobarbieFAQ and 1-888-256-0224 ©2015 Mattel. All Rights Reserved. ToyTalk and the ToyTalk logo are trademarks of ToyTalk. Apple, the Apple logo, and iPad are trademarks of Apple Inc., registered in the U.S. and other countries.
ToyTalk
MATTEL

Fig. 8.2 Hello Barbie (*Source* Mattel)

basis that doing so is an invasion of privacy. There is still a lot of uncertainty about how to manage human expectations in environments that feel as if they offer privacy, when actually they might not.

Hence we see that even large parts of a population may be willing to sacrifice their privacy if they see some benefit. Many cities around the world now contain cameras and systems in place for persistent surveillance, yet at least a number of countries seem to be willing to accept this. The use of unmanned aerial vehicles (UAVs) for the purpose of surveillance can also raise privacy issues. These systems have been used by individuals to spy on their neighbours. Regulations related to drone use vary from none at all to a complete prohibition on drone imports and use. The European Union, for example, has strong privacy protections in place which extend to data collected using a drone.

8.3.2 Usage of Private Data for Non-intended Purposes

While the availability of users' private data enables AI systems to perform better, there are also considerable risks associated with this data collection. One of the main issues is the usage of data for non-intended purposes. Users are often unaware how their data will be processed, used and even sold.

The success of programmatic advertisement demonstrates how personal data can be used to make people buy products. This specific form of marketing has in general been accepted by society, and people have developed protection mechanisms. We know that advertisements can only be trusted to some degree. But we have not yet developed such a critical awareness towards AI. We might not even be aware of manipulation taking place.

But there are serious forms of manipulation that use the exact same mechanisms that marketing experts applied so successfully: AI systems and social robots can be used to manipulate opinions on anything, including political views. Chapter 7 talks in more detail about the influence of AI on the public discourse. The Cambridge Analytica scandal in 2018 demonstrated how private data gathered through Facebook can be used in attempts to manipulate elections.

From a privacy perspective, the ability of modern AI to impersonate people has become a true danger. In 2016, Adobe demonstrated their VoCo system that can imitate the voice of any speaker after listening to approximately 20 min of conversation. But AI did not stop at speech. The manipulation of video, in particular the exchange of faces, has become a major issue that takes the problems of revenge porn to a new level. Spiteful ex-partners are able to mount the face of their previous love onto actors in porn movies and share them online. These so called "deep fakes" use neural networks to manipulate videos (see Fig. 8.3).

Another problem with the large scale collection of private data is that many people are unaware of the contracts they enter when signing up for various online services. The Terms and Conditions of social networks, for example, are not an easy read. The consequences, however, are not to be underestimated. As mentioned above, Facebook owns all the photos, messages and videos uploaded, and so does Google. And they do not shy away from selling this data to others. The General Data Protection Regulation (GDPR) in effect in the European Union since 2018 forced companies to seek consent from users before sharing their personal data with others. It should however also be noted that critics of the GDPR say that it leads to counterproductive results, by making small organisations and individuals shut down their websites for

Fig. 8.3 Former US president Barack Obama was used to showcase the power of deep fakes (*Source* BuzzFeedVideo)

fear of being heavily fined, while large companies can easily deal with the additional workload of complying with the new regulation.

Vulnerable Populations

Privacy violations and limitations can have different effects on vulnerable populations. Data generated by those receiving medical attention generates additional privacy concerns. This data, for instance, can be used not only to understand and treat an individual's affliction, but can also be used to infer things about the person's family genetics, physical and mental limitations, and perhaps even predict their death (White et al. 2012). Additional laws and regulations exist for health related data.

Artificially intelligent technologies foster dilemmas by generating data that would otherwise be private or using data in ways that were previously not possible. For example, gait (movement) information can be passively captured by camera observations of the person. This information could theoretically, in some cases, be used to predict an older adult's mortality. In a worst-case scenario, a private individual or business could passively observe large numbers of older adults, predicting their impending mortality, and then attempt to sell them funeral packages based on their data.

Children represent another vulnerable population. As mentioned above, data can be used to directly market products ranging from breakfast cereals to toys for children. Moreover, data generated when the child plays with an artificially intelligent toy can be transmitted to the company of origin and used for profiling, marketing, and advertising purposes. In general, parents tend to be more concerned about privacy when being observed with their children.

8.3.3 Auto Insurance Discrimination

Lack of privacy and the amount of data collected may lead to different rules being applied to different groups. Insurance companies, for example, value data in order to predict the cost of a policy and to assign premiums. Auto insurance companies use data to evaluate the driving behaviour of their drivers and to assign them to risk classes. AI use could lead to bias and discrimination here.

8.3.4 The Chinese Social Credit System

The Chinese government started work on a Social Credit System in 2014 that collects vast amounts of information about its citizens. While credit rating agencies have been operating for far longer, the Chinese government intends to extend the reach of its data collection far beyond what other organisations typically cover. The Chinese Social Credit System is already operational to the level of providing a financial credit score. In the future, it is intended to consider more private data, such as web browsing behaviour, and calculate how good a citizen is. This would have much further consequences than the denial of credit. Chinese citizens with a low score might be banned from flying and excluded from private schools, hotels and even careers.

Given the centralised authority of the Chinese government, its technical abilities and its plans for the future the Social Credit system could become the archetype for a tight control of the collection and use of private data. Its potential for use and abuse are still unclear. The opaque nature of the current system and the lack of a free press make it vulnerable to systematic bias.

8.4 Future Perspectives

From a technical perspective, it is relatively easy to apply the principles of programmatic advertisement to other application areas. Autonomous vehicles could, for example, make real-time bids for the lives of their passengers based on their social credit score in the face of an imminent crash. The passengers with the higher score would less likely to be harmed, while passengers with a lower social credit score would be exposed to greater risks of harm. The technical problems could be solved and only our societal discourse could prevent such a dystopian future.

It is also of concern that many have become ambivalent to how their private data is being shared, used and sold. Not long ago, people would have hesitated sharing a photo of themselves with the whole world. Today, social influencers flood the internet with revealing photos.

The future seems to be moving towards greater data collection. The value placed on privacy depends a great deal on the culturally and historical tendencies of the populace. Artificial intelligence expands the ways in which data is used, and may offer some predictive ability with respect to what is being collected. Future generations will need to decide the boundary that defines privacy and data use.

Discussion Questions:

- What information would you not like your digital assistant to have about you? Create a list.
- What parts of the population are most vulnerable to an invasion of their privacy by an AI? Explain your reasoning.
- Do you think a social credit system would have benefits? Discuss.

Further Reading:

- Michael J Quinn. *Ethics for the information age (7th edition)*. Addison-Wesley Publishing Company, 2017. ISBN 978-0134296548. URL http://www.worldcat.org/oclc/1014043739 (Chap. 5)
- Sare Baase and Timothy M. Henry. *A Gift of Fire Social, Legal, and Ethical Issues for Computing Technology (5th Edition)*. Prentice Hall PTR, 2017. ISBN 9780134615271. URL http://www.worldcat.org/oclc/1050275090 (Chap. 2).

Chapter 9
Application Areas of AI

> This section discusses several applications of AI to specific areas and illustrates the challenges and importance of ethics in these areas. While we cover autonomous vehicles and military uses of AI in separate chapters, here we discuss issues of AI for enhancement, healthcare and education.

9.1 Ethical Issues Related to AI Enhancement

It is starting to become possible to merge humans and machines. Robotic components that can replace biological anatomy are under active development. No ethical dilemma ensues when the use of robotic replacement parts is for restorative purposes. Robotic restoration takes place when a missing physical or cognitive capability is replaced with an equally functional mechanical or electronic capability. The most common example of robotic restoration is the use of robotic prosthetics. Enhancement, on the other hand, occurs when a physical or cognitive capability is replaced with an amplified or improved mechanical or electronic capability.

9.1.1 Restoration Versus Enhancement

It is not always clear when a restorative prosthetic becomes an enhancement. Technology is changing rapidly and prosthetics are becoming so advanced as to allow for significant increases in some abilities. Moreover, while some functions may only be restored, others may be enhanced. Typically, these distinctions do not inherently

C. Bartneck et al., *An Introduction to Ethics in Robotics and AI*,
SpringerBriefs in Ethics, https://doi.org/10.1007/978-3-030-51110-4_9

result in ethical dilemmas. Non-invasive mechanical and electrical devices, such as exo-skeletons, are also being developed to make people stronger, faster, or more capable in some ways. These systems may potentially be purchased by people for the sake of enhancement.

9.1.2 Enhancement for the Purpose of Competition

Ethical issues arise when we consider the use of enhancements for the purpose of job-related competition. Athletics offers an illustrative example. Athletes are often looking for competitive advantage over their rivals, and invasive or non-invasive augmentations may soon generate considerable advantage for those with the money to invest. Moreover, prosthetics are quickly becoming so capable that they offer a competitive advantage to some athletes. An ethical dilemma arises when these devices prevent fair competition or when they endanger the health of the athletes.

Note that the Superhuman Sports Society[1] promotes sports that explicitly invite the use of enhancements, but at least these enhancements are open and transparent. They even form an essential part of their games. Players without specific enhancements would not be able to play those games or would perform at a very low level. But we may ask how far will athletes go to be the best? In the 1990s, Goldman and Katz posed the following question to elite athletes: “Would you take a drug that would guarantee overwhelming success in your sport, but also cause you to die after five years?” The authors report that fifty percent stated yes (though other authors have since then disputed the results) (Goldman et al. 1987).

We can also consider cognitive versus physical enhancement. Amphetamine (methamphetamine in WWII) has been used to increase the wakefulness and energy of pilots and soldiers in combat. These enhancers were used by the United States Air Force until 2012. The United States Air Force still uses Modafinil as stimulant for combat pilots. Other militaries are suspected of using similar drugs.

Similarly, the number of college students using unprescribed Ritalin or Adderall on college campuses has tripled since 2008 (Desmon-Jhu 2016). Experts estimate that the total number of users may be more than 50% (Wilens et al. 2008). But is it ethical to use these so-called “Smart Drugs” to gain an academic advantage? Is this equivalent to doping in Olympic Sports?

More mechanical versions of cognitive enhancement are also becoming available. For example, transcranial magnetic stimulation of deep brain regions has been shown to improve cognitive performance. It may also soon be possible to make point-wise gene changes to improve biological, cognitive, or physical enhancements.

Although these technologies clearly have major benefits to society we must also be cautious to understand the potentially negative consequences. For the technologies described above, in particular, it is worth discussing and debating when an enhancement becomes a shortcut to excellence. Moreover, we must question whether or not

[1] http://superhuman-sports.org/.

the purchasing of cognitive aids cheapens success. At some point success may simply depend on one’s ability to purchase the necessary shortcuts. Less wealthy people would be disadvantaged in professional or academic success.

Another form of enhancement could be based on the brain-computer interfaces that are currently being developed. One of its prime applications is Neuroprosthetics, in which the neurons of a patient are connected to a computer which in turn controls a prosthetic device. Thinking about closing one’s hand will then result in the prosthetic hand closing.

9.2 Ethical Issues Related to Robots and Healthcare

Healthcare is another application of AI and robotics that raises ethical issues. Robots have been proposed for a wide variety of roles in healthcare including assisting older adults in assisted living, assisting with rehabilitation, surgery, and delivery. Currently robot-assisted surgery is the predominant application of robots within the healthcare industry. Robots are also being developed to deliver items in a hospital environment and for using ultraviolet light to disinfect hospital and surgical rooms.

9.3 Robots and Telemedicine

Robots have been suggested as an important method for performing telemedicine whereby doctors perform examinations and determine treatments of patients from a distance (see Fig. 9.1).

The use of robots for telemedicine offers both benefits and risks. This technology may afford a means for treating distantly located individuals that would otherwise only be able to see a doctor under extreme circumstances. Telemedicine may thus encourage patients to see the doctor more often. It may also decrease the cost of providing healthcare to rural populations. On the negative side, the use of telemedicine may result in and even encourage a substandard level of healthcare, when being used in an exaggerated way. It might also result in the misdiagnosis of certain ailments which are not easily evaluated remotely.

9.3.1 Older Adults and Social Isolation

Robots are also being introduced as a benefit to older adults to combat social isolation. Social isolation occurs for a variety of reasons such as children entering adulthood and leaving the home, friends and family ageing and passing away. Older adults that reside in nursing homes may feel increasingly isolated which can result in depression.

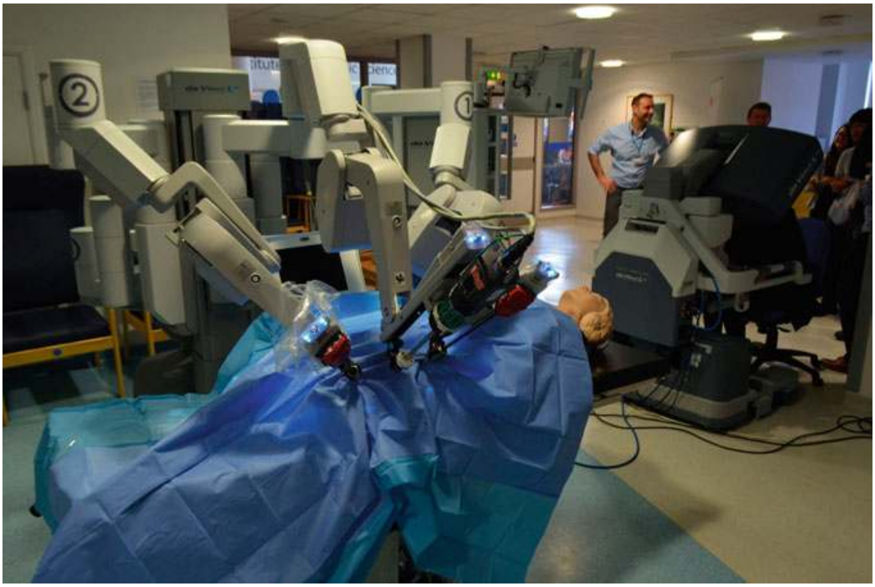


Fig. 9.1 da Vinci surgical system (*Source* Cmglee)

The United Kingdom acknowledged the societal scale problem of loneliness and appointed a dedicated minister in 2018.

Researchers have developed robots, such as Paro, in an attempt to reduce feelings of loneliness and social isolation in these older adults. Ethical concerns about the use of the Paro robot (see Fig. 9.2) have been raised (Calo et al. 2011; Sharkey and Sharkey 2012).

The main concerns are that patients with dementia may not realise that the robot is a robot, even if they are told (whatever the consequences may be). Moreover, the use of the robot may further increase actual social isolation by reducing the incentive of family members to visit. Yet, for this use case, many would argue that the benefits clearly outweigh the concerns (Abdi et al. 2018).

9.3.2 Nudging

Perhaps more controversial is the use of AI and robotics to provide encouraging nudges to push patients towards a particular behavioural outcome. Robotic weight-loss coaches, for example, have been proposed and developed that ask people about their eating habits and remind them to exercise (Kidd and Breazeal 2007). These robots are meant to help people stick to diets, but ethical concerns arise related to the issue of autonomy. Specifically, people should have the autonomy to choose how they want to live and not be subjected to the influence of an artificially intelligent

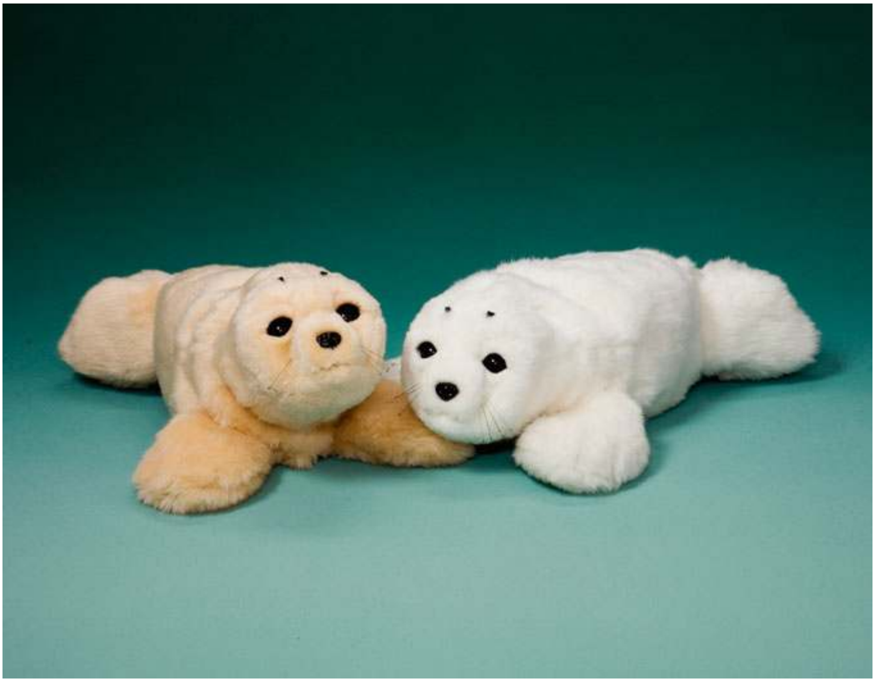


Fig. 9.2 Paro robot (*Source* National Institute of Advanced Industrial Science and Technology)

system. These systems also raise issues relate to psychological manipulation if their interactions are structured in a way that is known to be most influential. A variety of methods exist, such as the foot-in-the-door technique which could be used to manipulate a person.

9.3.3 Psychological Care

Recently, artificial systems have been proposed as a means for performing preliminary psychological evaluations. Ideally, these systems could be used to detect depression from online behaviour and gauge whether or not a treatment intervention is required. The use of such systems offers a clear benefit in that, by identifying individuals at risk, they may be able to prevent suicides or other negative outcomes (Kaste 2018). On the other hand, these systems still raise questions of autonomy and the potential development of nanny technologies which prevent humans from working through their own problems unfettered. Along a similar line of reasoning, virtual agents have been developed for interviewing Post Traumatic Stress Disorder (PTSD) suffers. Research has shown that individuals with PTSD are more likely to open up to a virtual agent than to a human therapist (Gonzalez 2017).

9.3.4 Exoskeletons

Looking at one final technology, exoskeletons have been developed to assist individuals with lower-limb disabilities. These systems are most often used for rehabilitation and training. Recently they have allowed paraplegic patients to stand and take small steps. Ethical issues arise when these systems are incorrectly viewed as a cure for disease rather than a tool.

9.3.5 Quality of Care

Overall, robotics and AI have the potential to revolutionise healthcare. These technologies may well provide lasting benefits in many facets of care ranging from surgery to diagnose disease. Certainly, a society must carefully analyse the benefits and costs of these systems. For instance, computer aided detection of cancer affects decisions in complex ways. Povyakalo et al. (2013) examined the quality of decisions that result when healthcare providers use computer-aids to detect cancer in mammograms. They found that the technology helped more novice mammogram readers but hindered more experienced readers. They note that this differential effect, even if subtle, may be clinically significant. The authors suggest that detection algorithms and protocols be developed that include the experience of the user in the type of decision support it provides. In 2019, a study on more than 9,400 women published by the Journal of the National Cancer Institute found that AI is overwhelmingly better in detecting pre-cancerous cells than human doctors (Hu et al. 2019).

9.4 Education

In education, AI systems and social robots have been used in a variety of contexts. Online courses are widely used. For example, the University of Phoenix is now (technically) one of the largest universities in the world—since hundreds of thousands of students are enrolled in their online courses. In such a highly digital learning environment, it is much easier to integrate AI that helps students not only with their administrative tasks, but also with their actual learning experiences.

9.4.1 AI in Educational Administrative Support

AI systems, such as Amelia from IPSoft, may one day advise students on their course selecting and provide general administrative support. This is not fundamentally different from other chatbot platforms such as Amazon's Lex, Microsoft's Conversation or Google's Chatbase. They all provide companies and organisations with tools to create their own chatbot that users can interact with on their respective websites or even on dedicated messaging platforms. While these bots may be able to provide some basic support, they do have to fall back to a human support agent when encountering questions that go beyond the knowledge stored in their databases.

Another form of supporting education with AI from an organisational perspective is plagiarism checking. In an age where students are able to copy and paste essays easily from material found online, it is increasingly important to check if the work submitted is truly the student's original work or just a slightly edited Wikipedia article. Students are of course aware of their teachers' ability to google the texts in their essays, and therefore are aware that they need to do better than a plain copy and paste. Good plagiarism checking software goes far beyond matching identical phrases and is able to detect similarities and approximations even search for patterns in the white space. Artificial intelligence is able to detect the similarities of text patterns and empowers teachers to quickly check student work against all major sources on the internet including previously submitted student contributions that were never openly published.

9.4.2 Teaching

AI systems have several advantages over human teachers that make them attractive for online learning. First, they are extremely scalable. Each student can work with a dedicated AI system which can adapt the teaching speed and difficulty to the students' individual needs. Second, such a system is available at any time, for an unconstrained duration at any location. Moreover, such an agent does not get tired and does have an endless supply of patience. Another advantage of AI teaching systems is that students might feel less embarrassed. Speaking a foreign language to a robot might be more comfortable for a novice speaker.

The autonomous teaching agents work best in constrained topics, such as math, in which good answers can be easily identified. Agents will fail in judging the beauty of a poem or appreciate the novelty in thought or expression. Section 2.4 discusses the limitations of AI in more detail.

Another teaching context in which robots and AI systems show promising results is teaching children with special needs, more specifically children with Autism Spectrum Disorder. The robots' limited expressivity combined with its repetitive behaviour (that is perceived by many as boring) is in this context actually a key advantage (Diehl et al. 2012). The use of robots for general purpose childcare is ethically questionable (Sharkey and Sharkey 2010).

9.4.3 Forecasting Students' Performance

Artificial Intelligence has been used to predict the dropout rate of students and their grades (Gorr et al. 1994; Moseley and Mead 2008). The goal is typically to provide students at risk with additional support, but also to carefully plan resource allocation. This is particularly important in the context of the United States' "No Child Left Behind" policy, which strongly encourages schools to minimise dropout rates and ensure good performance of most students.

There are, however, several ethical issues. First, the pressure applied under this policy can incentivise teachers and administrators to manipulate the scores of their students in order to meet the set targets. The Atlanta Public Schools cheating scandal is an example for such a misconduct (Fantz 2015).

Second, if the performance of a student is calculated prior to taking a course then both the student and the teacher might adapt to this score (Kolowich 2012). A student might, because of the prediction, not even try to perform well in the course. The student may deem their own effort as ineffective, exhibiting signs of learned helplessness. A teacher might also look at the list and decide that the students with the lowest predicted score are likely to drop out anyway and would not be worth putting extra effort into. These are two possible negative effects of such a forecasted performance, but both students and teacher might also decide to counteract the prediction. The student might work extra hard because he/she knows that this will be hard or the teacher might allocate extra time and resources to those students that are predicted to struggle. In any case, the consequences of using AI in predicting student performance should be discussed and monitored to avoid abuse and bias.

9.5 Sex Robots

One of the more controversial applications of AI technology is the design of robots for sexual purposes. Indeed, robots that engage in both sex and violence has been a trope in several recent hit films and TV shows such as *Ex Machina*, *Humans* and *Westworld* to name but three.

Those who argue against sex robots claim they will degrade people, especially women, and perpetrate harmful stereotypes of submissive females (Richardson 2016). It is certainly true that the vast majority of sex robots currently being produced are female. There are also concerns that giving people who lack social skills access to sex robots will cause them to not bother acquiring social skills. There are those who argue that sex is something humans should do with each other not machines. Some of these arguments are similar to arguments made against pornography and prostitution. There is even a Campaign Against Sex Robots[2] where more detail on these arguments can be found.

[2] https://campaignagainstsexrobots.org/.

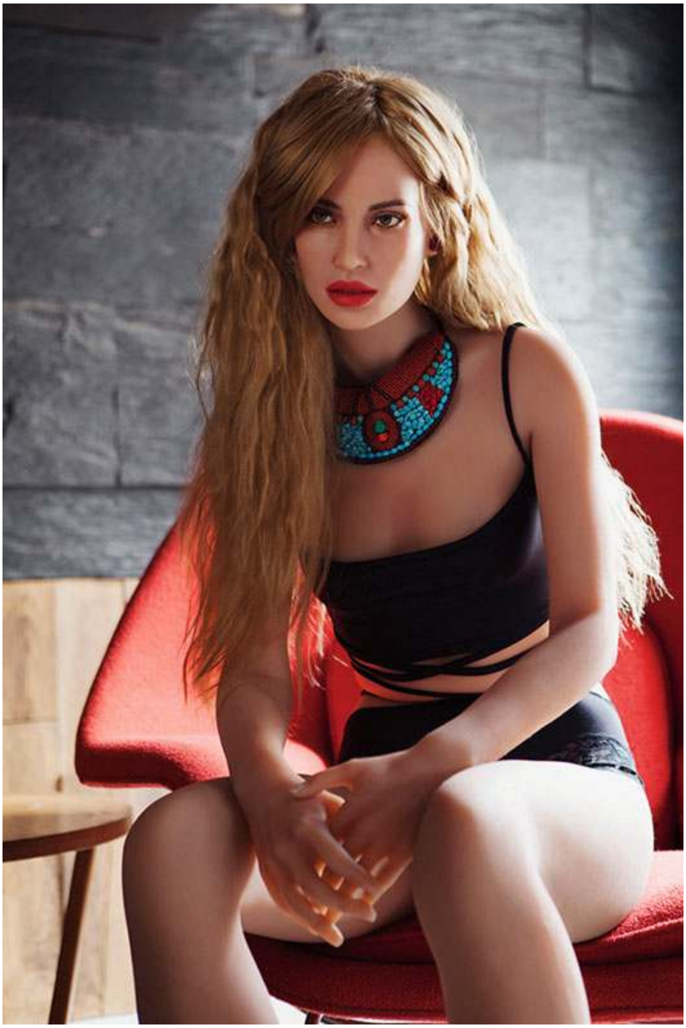


Fig. 9.3 A real doll (*Source* real doll)

Present day sex robots are little more than silicone dolls. These silicon dolls are, however, highly realistic and offer many customisation options (see Fig. 9.3).

Those who dismiss the arguments against sex robots argue there is no moral difference between a sex robot, a sex doll and a vibrator. The city of Houston modified its ordinance in 2018 (Ehrenkranz 2018) to ban the creation of a robotic brothel by changing its definition of Adult Arcades to include "anthropomorphic devices or objects that are utilized for entertainment with one or more persons". However, a sex robot with sophisticated AI could have some kind of relationship with a human. This would present the risk of unidirectional emotional bonding discussed in Sect. 7.3.

While most sex robots and sex dolls are female in form, male sex dolls are commercially available. Male sex robots with conversational AI may become available. Some think sex with a robot is in questionable taste. Others may argue that just because a thing is in poor taste is not necessarily a sufficient reason to ban it. The point being that this topic raises numerous ethical issues. For example, if a spouse has sex with a robot, does that count as infidelity? Would robot partners degrade human relationships (as depicted in *Humans*)? If the AI in a sex robot gets particularly advanced, will it get angry and turn on its human creators (an depicted in *Ex Machina*)? Is it acceptable (or possible) to "rape" and "kill" a sex robot which is a central theme of *Westworld*? Would this murder be morally equivalent to murder in a video game (Sparrow 2016)? At what point in their development should robots be given rights (Coeckelbergh 2010; Gunkel 2018) to protect them from human abuse? Coeckelbergh (2009) develops a methodology approach to evaluating roboethics questions related to person relations. At a point these questions become philosophical in nature (Floridi 2008).

Discussion Questions:

- If a drug would give you athletic triumph for five years then kill you, would you take it? Explain your reasoning.
- If predictive analytics suggested you should change your major, would you do so? Develop and discuss the costs and benefits of doing so.
- How would you feel about being tutored by an AI? Would you pay for this type of tutoring? Explain how much you would pay and why.

Further Reading:

- Patrick Lin, Keith Abney, and Ryan Jenkins. *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. Oxford University Press, 2017. ISBN 978-0190652951. URL http://www.worldcat.org/oclc/1011372036
- Illah Reza Nourbakhsh. *Robot futures*. MIT Press, 2013. ISBN 978- 026252 8320. URL http://www.worldcat.org/oclc/1061811057.

Chapter 10
Autonomous Vehicles

This chapter describes what autonomous vehicles are and how they work. We then discuss the associated risks and benefits not only from an ethical perspective, but also from a legal and environmental view. Core issues discussed are liability and safety of autonomous vehicles.

Some object to the term "autonomous vehicles." In philosophy and ethics, "autonomy" is a concept stating that human beings are to make decisions on their own. They should not be coerced by anyone, and they should be free to give themselves their own rules within reason (Greek: "autos" = self + "nomos" = law). This is not what AI or autonomous vehicles (AVs) could do. Some think a better term would be "highly automated vehicles" or "fully automated vehicles". In the context of AVs "autonomous" is used in the more standard robotic sense, which is simply the ability to operate for a protracted period of time without a human operator. However, since the term AV is well-established in the discussion, we will continue using it here. In the AV context, we need to distinguish between different levels or autonomy.

10.1 Levels of Autonomous Driving

There are several systems of classifying the levels of autonomous driving, in particular, one by the US Society of Automotive Engineers (SAE) and another by the German Association of the Automotive Industry (VDA). They differ, however, only in the details. The general idea behind these classifications goes as follows:

- **Level 0: No driving automation** This is a traditional vehicle without any automated functionality.

C. Bartneck et al., *An Introduction to Ethics in Robotics and AI*,
SpringerBriefs in Ethics, https://doi.org/10.1007/978-3-030-51110-4_10

- **Level 1: Driver assistance** The vehicle has one type of automated functionality. For example, braking automatically when encountering an obstacle.
- **Level 2: Partial driving automation** The vehicle can perform both braking and accelerating functions as well as changing lanes functions. However, the driver has to monitor the system at all times and be ready to take control whenever necessary. For example, all Tesla vehicles are officially considered Level 2 automation.
- **Level 3: Conditional driving automation** The driver does not need to monitor the system at all times. Under certain circumstances the system can work autonomously. The system gives the driver time (10 seconds for example) before handing back control. In 2018, the Audi A8 claimed to be the first car capable of Level 3 automation.
- **Level 4: High driving automation** The vehicle can perform all driving functions under standard circumstances. The driver is not required to take control under standard circumstances. Non-standard conditions would include inclement weather.
- **Level 5: Full driving automation** The vehicle can perform all driving functions in all circumstances. In the German classification, this is labelled as "no driver", making the car completely autonomous.

10.2 Current Situation

As of 2019 there are no commercially available autonomous vehicles beyond level 2 or level 3 automation. It is interesting to note that a number of new companies are challenging the traditional manufacturers, with some of them being on the market already (Tesla), and some of them are testing and collecting data while preparing for market entry at a later time (Waymo see Fig. 10.1). Test driving is, on the other hand, widespread already. Nevada gave Google a permit to test AVs in 2009. In Europe, the Netherlands, Germany and the UK permit AV testing, A number of US federal states have passed or changed regulations to allow autonomous test driving under certain conditions.

Testing of AVs in the US started in 1995 when an experimental autonomous vehicle developed by Carnegie Mellon University drove from Pittsburgh to San Diego. Rideshare platforms like Uber of Lyft are developing and testing autonomous taxis. Singapore completed a test on autonomous taxis in 2018. Japan is planning to introduce autonomous taxis on a large scale for the Tokyo 2020 Olympics. Moreover, in a number of places, autonomous buses are employed, usually in controlled environments.

10.3 Ethical Benefits of AVs

According to the available studies, AVs could bring about a number of safety benefits, especially by avoiding accidents and fatal crashes. It is estimated that around 90–95% of all automobile accidents are chiefly caused by human mistakes (Crew 2015).

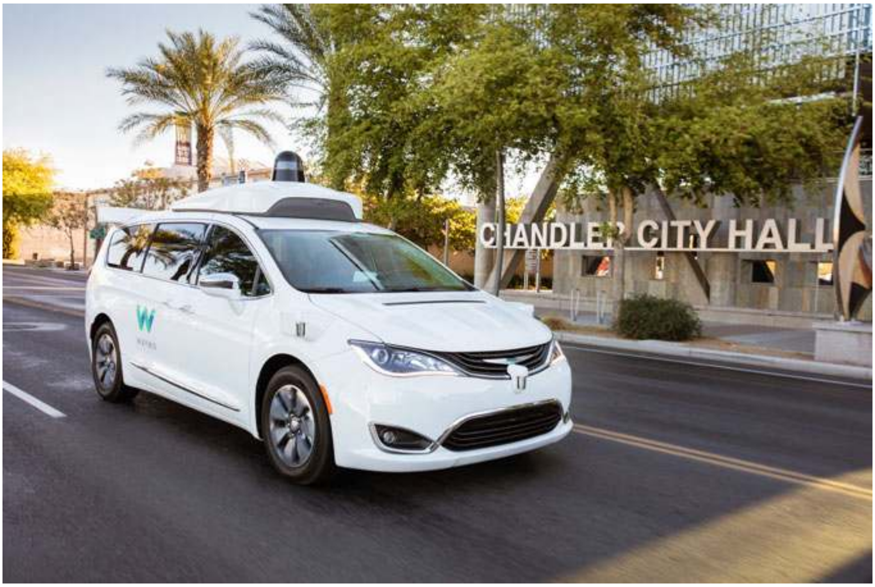


Fig. 10.1 Waymo's fully self-driving Chrysler Pacifica Hybrid minivan on public roads (*Source* Waymo)

This raises the question of whether or not a society might be ethically obligated to legally permit autonomous vehicles.

Still, some argue that the presence of AVs may indirectly induce certain types of accidents when the AVs strictly follow the law: For example, following the Atlanta highway speed limit of 55 mph, when the speed of the average human driver might rather be 70 mph, will likely cause people to navigate around the autonomous vehicle, possibly resulting in a greater number of "human-caused" accidents. However, the autonomous cars cannot be blamed for these type of accidents since human drivers did not observe the rules of traffic in the first place.

10.4 Accidents with AVs

There have already been some accidents with autonomous vehicles. Even though the number is small, they have generated a lot of media attention. One of the first was an accident with a Tesla in 2016, where the vehicle did not recognise a truck, killing the driver. Sensory equipment has been improved since then (Hawkins 2018).

In March 2018, an Uber Volvo fatally hit a woman crossing the road in Tempe (Arizona), an accident which probably no human driver could have avoided. The car's emergency braking system was however deactivated on purpose to avoid false positives, and the driver was not acting appropriately (O'Kane 2018).

Another (fatal) accident happened with Tesla cars in March 2018 in Mountain View/California, and in Laguna Beach/California in May 2018. In both cases, it turned out that the respective driver ignored several warnings by the "autopilot" system.

These (and similar) cases show we need clear rules about how AVs are to be used by customers. Drivers should not sit behind the wheel playing games or reading their email. Companies should be required to communicate these rules in adequate ways.

10.5 Ethical Guidelines for AVs

Besides detailed regulations that have been passed in a number of countries, a broader ethics code for AVs has been adopted. In July 2016, the German Federal Minister of Transport and Digital Infrastructure appointed a national ethics committee for automated and connected driving. The committee was composed of 14 members, among them professors of law, ethics and engineering, as well as representatives from automotive companies and consumer organisations. The chairman was a former judge of the German Federal Constitutional Court. In addition, hearings with additional experts from technical, legal and ethical disciplines were conducted, as well as driving tests with several AVs. In June 2017, the committee presented 20 ethical guidelines for AVs (Federal Ministry of Transportation and Digital Infrastructure 2017). Many of these will have a significant direct or indirect bearing on the car industry and their ethics policies, both for Germany as well as the EU (Luetge 2017).

There are some researchers that feel that caution is necessary when discussing the ethical benefits of autonomous vehicles (Brooks 2017; Marshall 2018) and risks associated with autonomous vehicles (Cummings 2017). One issue is that the data of miles driven in autonomous mode might be skewed in terms of ideal driving conditions, that the behaviour of autonomous vehicles differs from certain 'unwritten' norms of human drivers (Surden and Williams 2016), and that people may place themselves at greater risk because of the presence of autonomous vehicles (Rothenbücher et al. 2016). The RAND corporation suggested in 2016 that an AV must drive 275 million miles without a fatality to prove that they are safe (Kalra and Paddock 2016). RAND researchers later however also stated that AVs should be deployed as soon as possible in order to save lives (Marshall 2017).

10.6 Ethical Questions in AVs

At least some of the accidents indicate that it may be ethically questionable to call a certain mode "autopilot" (which Tesla does), suggesting false connotations, since the driver of a Tesla car (which is officially Level 2) is required to keep the hands on the wheel at all times. Just stating that this is a matter of one's own responsibility is insufficient—as with many other cases from different industries. In the 1970s, for example, a massive scandal erupted for the Nestle company, as people believed it did not give adequate rules for mothers how to use their baby milk products. In the not-so-distant future, AVs might be introduced on a massive scale. Companies have a responsibility lay out drivers' responsibilities very clearly.

In addition, there are a number of other ethical questions usually associated with AVs.

10.6.1 Accountability and Liability

The question of liability will be critical. According to the 1968 Vienna Convention on Road Traffic (Vienna Convention on Road Traffic 1968, updated 2014, effective since 2016 (United Nations 1968), the driver is the one responsible for their car. However, in cases where an autonomous system is in control, this responsibility does not make much sense. The German Ethics Code therefore states that in these cases, the liability has to be turned over to the car manufacturer and the company operating or developing the software. Liability in autonomous driving will become a case of product liability and it will also require monitoring devices to be built into AVs. Future AV driving recorders might be similar to the flight recorders used on contemporary aircraft.

10.6.2 Situations of Unavoidable Accidents

Much of the literature on ethics of AVs revolves around situations similar to the "trolley problem" where an accident is unavoidable and an AV has to choose between two evils (Bonnefon et al. 2016; Lin 2016). Should the car choose to swerve to avoid hitting four people and hit one person instead? Does it matter if the people are children, or older adults? Should it hit someone crossing the street at a red light rather than someone observing the traffic rules (see Fig. 10.2)?

How frequent these situations will be in practice is controversial. Some have argued that these situations will be quite rare, given that the conditions for them are extremely narrow. But if we accept their possibility for the moment, then one issue is to avoid creating a machine or an algorithm that selects targets according to personal characteristics (such as age or gender, which the German Ethics Code prohibits, or Social Credit Score), but still allowing a company to program code that reduces the overall number of fatalities or injuries. This is a complex problem, which is usually not as simple as selecting one of several targets to be killed with a guarantee. For example, there will probably be different probabilities for injuries or casualties of different targets. This may result in complicated situations in which it would not be ethical to forego the opportunity to reduce the overall damage to persons. We should make it clear that current autonomous vehicles do not engage in trolley problem

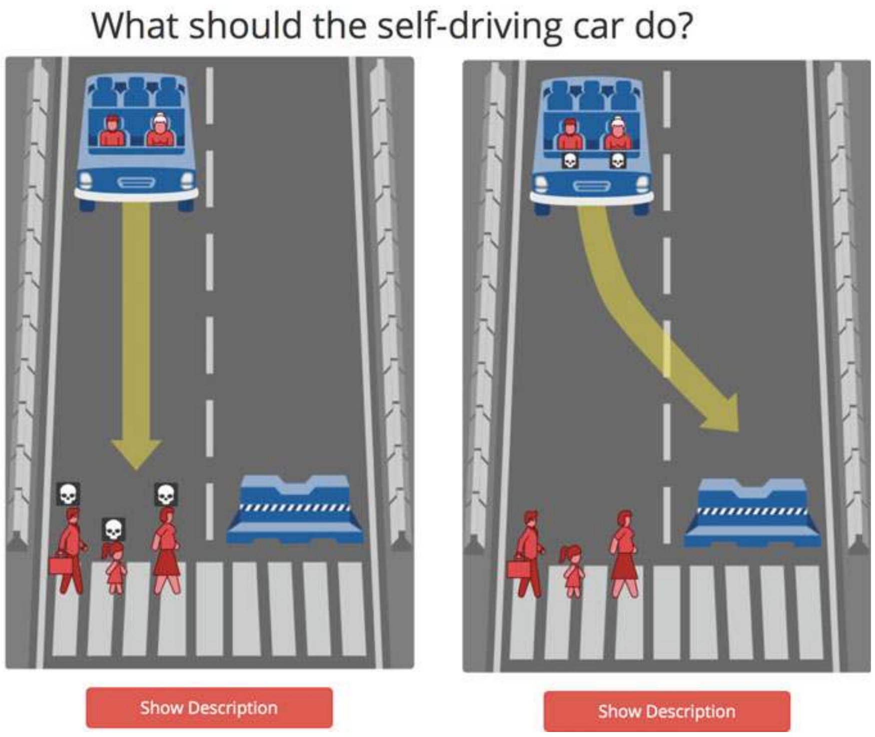


Fig. 10.2 Example question from the Moral Machine experiment that confronted people with trolley problems (*Source* MIT)

calculations. These vehicles attempt to avoid all obstacles and make no attempt to identify the obstacle, although in some cases they avoid obstacles based on size.

Cars do already know the number of passengers since they feature a seat belt warning system. Car to car wireless communication is also already available. In the case of an imminent crash, it would be technically feasible for the two cars to negotiate driving behaviour, similar to the Traffic Collision Avoidance System used already today in air planes. The car with more passengers might be given priority. While this technical possibility exists, it will be up to society to decide if such a system would be desirable.

However, what the German code explicitly does not say is that individual victims in different scenarios are allowed to be offset against each other. To some extent, this is the lesson of the German "Luftsicherheitsgesetz" (Aviation Security Act) being ruled unconstitutional by the German Federal Constitutional Court in 2006. The Aviation Security Act would have allowed to shooting down of hijacked aircraft that were thought to be used as weapons. In that case, individually known subjects would have been sacrificed for the sake of others. In the case of an anonymous program, however, no victims are known individually in advance. Rather, it is an abstract guideline, the exact consequences of which cannot be foreseen, and which reduces the overall risk for all people affected by it. In this way, the risk could be regarded as similar to the risk that comes with vaccination.

Finally, parties not involved in the traffic situation with an AV should not be sacrificed. This implies that an algorithm should not unconditionally save the driver of an AV. However, as the German Ethics Code states, the driver should not come last either: after all, who would want to buy such a car?

10.6.3 Privacy Issues

The problem of privacy, which has come to the forefront especially with developments like the European General Data Protection Regulation, effective 2018 (see Sect. 8), is also relevant for AVs. These cars collect massive amounts of data every second they move, which is essential for evaluating their performance and improving their safety. On the other hand, such data may be very personal, tracking the passengers' locations and the behaviour of the driver. Such data might be used for purposes for which they were not intended. Ultimately, starting the engine of an AV will probably imply that the driver (and/or the passengers) accept the terms of conditions of the car manufacturer who will require access to such data.

It should also be noted that this problem is seen from different perspectives in different regions around the globe, especially considering the US or China, where data collection is seen as being less problematic than, for example, in Germany or other EU countries.

Autonomous vehicles generate data as they move from one location to another. This data may include observations of the individuals inside the car. Moreover, observations can be used to evaluate whether or not a person should be allowed to drive the car. Technology could be used to prevent people with suspended license or drinking problems from driving. For an autonomous vehicle, the vehicle itself may determine if the person's behaviour indicates intoxication and decide whether or not to allow the person to drive. Perhaps more of a concern, is when an autonomous vehicle uses data about the person's emotional state, mood, or personality to determine if their driving privileges should be curtailed.

10.6.4 Security

Security of AVs is an important issue, since hackers might breach the security of cars or their systems and cause substantial harm. In 2015, for example, hackers were able to remotely control a Jeep Cherokee while driving (Greenberg 2015). Cars might be reprogrammed to deliberately crash, even on a massive scale. Therefore, state-of-the art security technology will have to be involved in order to make AVs as secure as possible.

10.6.5 Appropriate Design of Human-Machine Interface

The interface between a human driver and the AV is very important, since in a traffic situation, time is highly critical (Carsten and Martens 2018). It must be clear at all times who is in control, the handover procedure must be clearly defined, and the logging of behaviour and car control must be programmed in an appropriate way.

10.6.6 Machine Learning

AV software will certainly use machine learning in an offline manner. The systems will learn from traffic scenarios they encounter, from accidents or from the data generated by other vehicles. Vehicles, however, are not dynamically learning who and what to avoid while they are driving. Moreover, learning typically occurs on the level of an entire fleet of cars. And even here, learning must be robust to an extent where small mistakes (or external manipulations) cannot cause large-scale negative effects. Vehicles that learn on their own, might become quite erratic and unpredictable.

10.6.7 Manually Overruling the System?

Finally, one unresolved problem is whether a level 4 or 5 AV should be programmed in such a way as to allow the driver to overrule the autonomous system at any point. Engineers regularly reject this possibility stating that this feature would make a system much more vulnerable than it could be. Waymo even argued that its car would have avoided a collision in 2018 with a motorcyclist if the driver had not taken back control (Laris 2018).

Others argue that to deny overruling would jeopardise acceptance of AVs. At least in the current situation where people are not yet familiar with the technology, it might be better to allow for overruling, making a driver feel more comfortable. But this may be reviewed at some point in the future.

10.6.8 Possible Ethical Questions in Future Scenarios

While the ethical questions mentioned above are relevant today, there are some ethical aspects of AVs which are frequently discussed in public, even though they may be only relevant in the distant future. One of these is whether the introduction of AVs on a massive scale might lead to a critical centralisation of power in the control centres of this technology, which in turn might lead to a total surveillance of citizens and/or a deliberate or accidental misuse. Certainly, there are some scenarios here that should be kept in mind, even if a Big Brother scenario is unrealistic since it neglects the competition among different companies.

Another dystopian outcome that is frequently mentioned is whether AVs might become mandatory at some point in the future, forbidding the use of non-autonomous vehicles at all (Sparrow and Howard 2017). We believe that at this point, it is much too early to decide on this. There should certainly be a public discussion about whether such a step should be taken. But, right now, we are technically far from its implementation, and will rather, at least for the next decades, have to deal with a situation with mixed traffic in which autonomous vehicles will mix with traditional ones. Studies predict that this situation will already substantially prevent many accidents and save a lot of lives (Crew 2015).

Discussion Questions:

- On what level of "autonomy" (according to SAE and VDA) would you be comfortable to drive alone? What if you were with your child? Explain your reasoning.
- According to the Vienna Convention, who is responsible for a car? Is this law appropriate for AVs? Why? Why not?
- Do you think that young and old people should be treated differently by an AV's algorithm? Why? Why not?

Further Reading:

- Jean-Francois Bonnefon, Azim Shariff, and Iyad Rahwan. The social dilemma of autonomous vehicles. *Science*, 352(6293):1573–1576, 2016. ISSN 0036-8075. Doi: 10.1126/science.aaf2654. URL https://doi.org/10.1126/science.aaf2654
- Noah J. Goodall. Can you program ethics into a self-driving car? *IEEE Spectrum*, 53(6):28–58, 2016. ISSN 0018-9235. Doi: 10.1109/ MSPEC.2016.7473149. URL https://doi.org/10.1109/MSPEC.2016.7473149
- Patrick Lin. Why ethics matters for autonomous cars. In Markus Maurer, J Christian Gerdes, Barbara Lenz, Hermann Winner, et al., editors, *Autonomous driving*, pages 70–85. Springer, Berlin, Heidelberg, 2016. ISBN 978-3-662-48845-4. Doi: 10.1007/978-3-662-48847-8_4. URL https://doi.org/10.1007/978-3-662-48847-8_4
- Christoph Luetge. The German ethics code for automated and connected driving. *Philosophy & Technology*, 30(4):547–558, Dec 2017. ISSN 2210-5441. Doi: 10.1007/s13347-017-0284-0. URL https://doi.org/10.1007/s13347-017-0284-0.

Chapter 11
Military Uses of AI

In this chapter we discuss the controversial topic of autonomous weapons systems. We define terms, present elements and perspectives related to the debate surrounding autonomous weapons systems (AWS) and present the main arguments for and against the use of such systems.

War is inherently controversial. Many military uses of AI and robotics are likewise contentious. Perhaps the most controversial aspect of this topic is the development and use of lethal autonomous weapons systems capable of autonomously making life and death decisions regarding human targets. Some argue that cruise missiles are a form of lethal autonomous weapons system. Systems such as the Patriot missile system, AEGIS naval weapons system, Phalanx weapons system, and Israeli Harpy weapons system are examples of lethal autonomous weapons systems currently in use. The Patriot system, AEGIS, and the Phalanx system are generally considered defensive weapons (Fig. 11.1).

The Harpy is an offensive fire-and-forget weapon that targets enemy air warfare radar systems. It is worth noting that the term military robot includes many non-lethal applications. For example, autonomous robots may be used for mine clearing, explosive ordnance disposal, command and control, reconnaissance, intelligence, mobile network nodes, rescue missions, supply and resupply missions, and support operations. Debates against military robots may vary with respect to the role of the robot. The aim of this chapter is to present an objective account of the most commonly presented arguments for and against the use of AWS in war.

C. Bartneck et al., *An Introduction to Ethics in Robotics and AI*,
SpringerBriefs in Ethics, https://doi.org/10.1007/978-3-030-51110-4_11

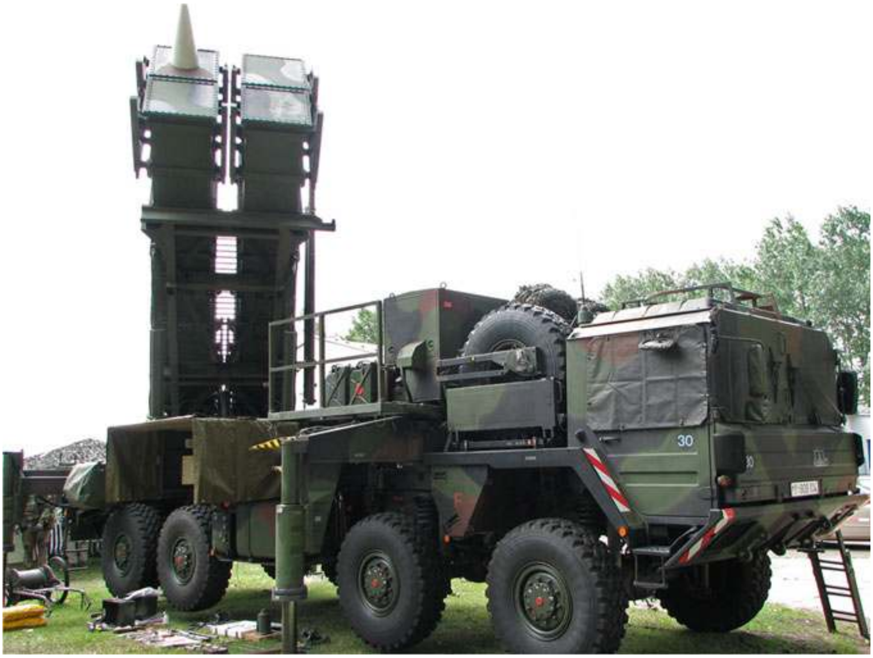


Fig. 11.1 MIM-104 Patriot (*Source* Darkone)

11.1 Definitions

We begin by defining some common terms.

Autonomous In AI and robotics autonomy simply means the ability to function without a human operator for a protracted period of time (Bekey 2005). Robots may have autonomy over the immediate decision that they make but generally do not have autonomy over their choice of goals. There is some controversy as to what "autonomous" means for lethal weapons systems.

Lethal and Harmful autonomy A weapon can be said to be "autonomous" in the "critical functions of targeting" if it can do one or more of following without a human operator. If the weapon can decide what classes of object it will engage, then it would be autonomous in terms of defining its targets. No current AWS has this capability. If a weapon can use sensors to select a target without a human operator, it can be said to have autonomy in the selection function of targeting. Many existing weapons can select targets without a human operator. If a weapon can fire on a target without a human operator, it can be said to have autonomy in the engage function of targeting. Many existing weapons can engage already selected targets autonomously. For example, the Patriot anti-missile system can select targets autonomously but by design requires a human operator to hit a confirm button to launch a missile. Once the missile is launched, it can hit its target without a human operator. Given the speeds involved, human control of a Patriot missile is not be possible.

Non-Lethal Autonomy An AWS may have "autonomy" in many other functions. It might be able to take-off and land autonomously and it might be able to navigate autonomously. However, this non-lethal "autonomy" is generally not regarded as morally controversial.

Killer robots Autonomous weapons are often called "killer robots" in mass media reports. Some object to the use of the term. Lokhorst and van den Hoven describe the phrase as an "insidious rhetorical trick" (Lokhorst and Van Den Hoven 2012). However, this is favoured by the "Campaign to Stop Killer Robots".[1] This is umbrella group of human rights organisations seeking an international ban on lethal autonomous weapons systems.

11.2 The Use of Autonomous Weapons Systems

Arguments can and are made against the use of an autonomous weapons system. Generally these arguments focus on the following issues.

11.2.1 Discrimination

Proponents typically concede that machines cannot, in general, discriminate as well as humans. However in some particular cases they can discriminate better than humans. For example, Identification Friend or Foe (IFF) technology sends a challenge message to an unindentified object in the sky which the object must answer or risk being shot down. Typically in air war, contested airspace is known to civilian air traffic control and neutral aircraft will not enter it. However, in 2014, a civilian airliner, Malaysia Airlines flight MH 17 en route from Amsterdam to Kuala Lumpur was shot down by a Russian Surface to Air Missile (SAM) operated by Russian secessionists in the Eastern Ukraine. This SAM system was human-operated and not equipped with IFF. Some have observed that a more advanced system would have known the target was a civilian airliner.

Proponents also note that vision systems are continuously improving. Advancing technology has dramatically improved the ability of vision, auditory, LIDAR and infra-red systems which are quickly reaching parity with humans in terms of object discrimination. A possible ethical dilemma may be approaching if an autonomous system demonstrates clear superiority to humans in terms of targeting. We may be ethically obligated to consider their use. Even so, it remains difficult for machines to distinguish between different types of behavior such as acting peaceful or fighting in a conflict.

[1] http://stopkillerrobots.org.

Opponents note that it has been frequently claimed that AWS cannot discriminate between combatants and non-combatants. Noel Sharkey, a leading campaigner against AWS, has wondered if a robot could discriminate between a child holding an ice-cream cone and a young adult holding a gun (Sharkey 2010).

11.2.2 Proportionality

Proponents, on the other hand, state that "excessive" is a relative concept which is not well-defined in International Humanitarian Law (IHL). Enemark makes the point that politicians generally do not advertise their proportionality calculations (Enemark 2013). Situations in which intelligence reveals the location of a high value target demand a decision.

Opponents claim that AWS cannot calculate proportionality (Braun and Brunstetter 2013). Proportionality is the ability to decide how much collateral damage is acceptable when attacking a military target. The standard is that "collateral damage" must not be "excessive" compared to the concrete military advantage gained. Proportionality calculations typically attempt to estimate the number of civilians that may be killed versus the military necessity of the target. Generating such calculations often involves input from a variety of experts including lawyers. It is difficult to imagine how an AWS could successfully complete such a calculation.

11.2.3 Responsibility

Opponents of AWS argue that machines cannot be held morally responsible. They then argue that this is a reason to ban AWS. It is indeed hard to imagine how a machine can be assigned moral responsibility. However those defending the use of AWS are inclined to assign moral responsibility for the actions of the machine to those that design, build and configure it. Thus those humans deploying an AWS can be held responsible for its actions (Arkin 2008). This raises the "problem of many hands" (Thompson 1980) in which the involvement of many agents in a bad outcome makes it unclear where responsibility lies. Clearly if an incident were to occur an investigation would result to determine fault. Legally it is easier to hold the collective entity responsible. There is a concept of "strict liability" in law that could be used to assign responsibility to the state that operates the weapon in an AWS regulation. We discuss liability with the more specific example of an mistargeting by an autonomous weapon in Sect. 5.2.

Opponents also argue that, unlike a human, an AWS cannot be held responsible for its actions or decisions. While machines can be grounded for performance errors there is no true way to punish these systems in a metaphysical sense. Moreover, it may not be just to punish the commanders of these systems if they utilize automatic targeting.

11.3 Regulations Governing an AWS

States at the UN agree that an AWS must be used in compliance with existing IHL. The Convention on Certain Conventional Weapons is generally considered the appropriate forum to discuss AWS regulations. These regulating bodies require meaningful human control over the AWS. In particular, this means the following:

1. an AWS must be able to distinguish between combatants and non-combatants;
2. an AWS must be able to calculate proportionality;
3. an AWS must comply with the principle of command responsibility.

11.4 Ethical Arguments for and Against AI for Military Purposes

11.4.1 Arguments in Favour

In IHL the doctrine of military necessity permits belligerents to do harm during the conduct of a war. Moreover, just war theory states that, although war is terrible, there are situations in which not conducting a war may be an ethically and morally worse option (International Committee of the Red Cross 2015). For example, war may be justifiable to prevent atrocities. The purpose of just war theory is to create criteria that ensures that war is morally justifiable. Just war theory includes criteria for (1) going to war (jus ad bellum) and (2) conducting war (jus in bello). The criteria for going to war include: just cause, comparative justice, competent authority, right intention, probability of success, last resort, and proportionality. The criteria for conducting war include: distinction, proportionality, military necessity, fair treatment of prisoners of war, and not using means and methods of warfare that are prohibited. Examples of prohibited means of warfare include chemical and biological weapons. Examples of prohibited means include mass rape and forcing prisoners of war to fight against their own side. The overall intent of IHL is to protect the rights of the victims of war. This entails rules that minimise civilian harm.

With respect to the use of AI and robots in warfare, some have argued that AMS may reduce civilian casualties (Arkin 2010). Unlike humans, artificially intelligent robots lack emotions and thus acts of vengeance and emotion-driven atrocities are less likely to occur at the hands of a robot. In fact, it may be the case that robots can be constructed to obey the rules of engagement, disobeying commands to violate civilian and enemy combatant rights (Arkin 2008). If nothing else, units being observed by a military robot may be less inclined to commit such atrocities. If, in fact, robots can be used to prevent atrocities and ensure the minimisation of civilian casualties, then military leaders may have an ethical obligation to use such robots. For, to not use such robots, condemns a greater number of civilians to die in a morally justified war. Moreover, an AWS may be capable of non-lethal offensive action where human units must use lethal force.

Other argue that AI and military robots are necessary for defensive purposes. Some research has shown that in certain circumstances, such as aerial combat, autonomous systems have clear advantages over human systems (Ernest et al. 2016). Hence, sending humans to fight an AWS is unlikely to succeed and may result in substantial casualties. In this situation, leaders have an ethical obligation to reduce their own casualties even if this means developing AWS for their own purposes.

11.4.2 Arguments Against

It has been claimed that the advent of artificial intelligence technologies for military use could lead to an arms race between nations. Vladimir Putin, the President of the Russian Federation, said in 2017 that "the nation that becomes the leader in AI will rule the world." (James 2017). China has similarly increased spending on AI (Herman 2018) and the United States has long made the development of AI for defense purposes a priority (Department of Defense 2012). Experts generally agree that AWS will generate a clear and important military advantage (Adams 2001). Relatedly, some argue that use of an AWS is unfair in that such weapons do not result in equal risk to all combatants.

Researchers also note that the possession and use of autonomous weapons systems may actually instigate wars because the human cost of war is reduced. There is some evidence for this claim based on targeting killings in Iraq by the United States. The transition in Iraq from human piloted missions to unmanned aerial vehicles resulted in a dramatic increase in the number of targeting missions (Singer 2009). This evidence, although important, should not discount the political and technological factors that may also have contributed to the increase in targeted killings.

Perhaps the most philosophically interesting argument levelled against the use of AWS is the dignity argument claiming that "death by algorithm" is the ultimate indignity. In its more complex forms, the argument holds that there is a fundamental human right not to be killed by a machine. From this perspective, human dignity, which is even more fundamental than the right to life, demands that a decision to take human life requires consideration of the circumstances by a human being (Arkin et al. 2012; Heyns 2016). A related claim is that meaningful human control of an autonomous weapon requires that a human must approve the target and be engaged at the moment of combat.

11.5 Conclusion

To conclude we have sought to present, relevant definitions associated with autonomous weapons systems, the ideas behind the regulations that govern these systems, and the arguments for and against their use. An AWS cannot be lawfully used for genocide and massacre of civilians because existing humanitarian law already prohibits such acts. It is important to note that AWS are already regulated and must be used in accordance with existing international law. It is important to keep in mind that these systems are changing. As they do, they may raise new and important ethical issues that should be discussed within and between nations. Further, in much the same way that commanders are responsible for the actions of their soldiers, commanders are also responsible for the actions of their AWS.

Discussion sQuestions:

- Is the use of AWS in military conflicts justified? Explain.
- What limits or conditions would you set before an AWS could be used? List a set of conditions.
- Should there be new IHL for AWS? Discuss.

Further Reading:

- Ronald Arkin. *Governing lethal behavior in autonomous robots*. Chapman and Hall/CRC, 2009. ISBN 978-1420085945. URL http://www.worldcat.org/oclc/933597288
- Peter Warren Singer. *Wired for war: The robotics revolution and conflict in the twenty-first century*. Penguin, 2009. ISBN 1594201986. URL http://www.worldcat.org/oclc/958145424
- Paul Scharre. *Army of none: Autonomous weapons and the future of war*. WW Norton & Company, 2018.

Chapter 12
Ethics in AI and Robotics: A Strategic Challenge

The recent success of AI and robotics has massively increased the international awareness of and the interest in these topics as a factor of economic competitiveness. This concerns not only businesses, but also regional and national governments. There are many claims about the potential for AI to create new service and product innovation. Such claims include the benefits of AI for healthcare such as improved diagnosis or therapy; for transport due to improved efficiency; for energy based on more accurate predictions of energy consumption; or ease of computer use with more natural user interfaces as, for example, in the case of speech understanding, gesture and face recognition and automatic translation.

In general, many smart or intelligent technologies have been considered a major driver of innovation (Lee and Trimi 2018; Makridakis 2017), and also an important source of knowledge for innovation (Fischer and Fröhlich 2013). As a result of these promises, there is today a plethora of regional, national, and even supranational strategies and policy papers that aim at maximising the benefits of AI for their citizens. Examples of national strategies range from Canada to Mexico, Japan or India. Regional strategies have been developed in Bavaria and in the Northern Baltic countries. Supranational AI strategies, or at least joint studies and papers are the subject of work in the OECD and the United Nations International Telecommunication Union (ITU).

In early 2019, a broad range of policy papers (Agrawal et al. 2019) and marketing studies from consulting companies have been published. Many of these make the case for the innovation potential and economic benefits of AI (Li et al. 2017; Seifert et al. 2018). Governments around the world have responded to the massive increase in AI applications, but also to an even greater number of predictions of future AI applications and their societal benefits. As early as 2017, the Canadian government published a pan-Canadian AI strategy. It was followed by developed countries traditionally interested in the creating information technology such as Japan, Singapore,

C. Bartneck et al., *An Introduction to Ethics in Robotics and AI*,
SpringerBriefs in Ethics, https://doi.org/10.1007/978-3-030-51110-4_12

Finland and China. By mid-2018 the European Commission published its Communication on Artificial Intelligence, thus effectively motivating its member states to draft strategies for AI. In December 2018, the EU presented its plan for AI with more concrete actions from research to development, investments in AI, ensuring training and education and a proper computing infrastructure.

AI strategies around the world mostly follow a general model that addresses the **actors** in the AI and robotics environment, such as professionals, society, research organisations, companies and government. These groups require support through **infrastructure**, data, staff, finance, and information to productively create an environment conducive to the deployment of AI solutions. The aim is to create AI **solutions**, practices and improved benchmarking within industry. To support this, government strategies focus on a range of **processes** ranging from cooperation between these actors, on improving research excellence, staff training, and regulation. In addition, many strategies emphasise the need for an elaborate societal discourse and an ethical approach to AI.

It is not only the promises of increased competitiveness or new applications, which are driving the development of national strategies for AI and robotics. On the contrary, it seems that the public discussion in many countries is more focused on the potential damages that AI may induce, for example in the labour market, but also regarding human autonomy, privacy, and even the very future of society. There are two main streams that present one or another version of AI—and often robotics—dystopia. From an economic perspective, authors like Ford (2015) studied the potential impact of AI and robotics on (human) work. He predicts massive job losses in many sectors that have seemed to be immune to automation for a long time. The argument here is that new AI technology is now capable of replacing much more of the work for which human intelligence was required to date. This includes, for example, medical diagnostic knowledge, expert knowledge from tax advisory or the legal knowledge of lawyers. The second line of dystopian publications stems from more journalistic accounts of the potential future of AI. In several cases, these publications draw an image of our future in which AI overlords threaten humanity while others are more careful predictions about losing our privacy and autonomy (Bartlett 2018).

These often rather pessimistic predictions about the impact of AI have been quite successful in terms of their influence on the broader public. It is therefore unsurprising that policy makers around the world include the potential damage created by AI and robotics in their considerations and discussions. For example, the German AI strategy explicitly calls for a broad discussion of AI's societal impacts. In addition, it aims to support a continued discussion between politics, science, industry, and society. Several national policies emphasise the need to continuously monitor and study the impact of AI technology on labour and society. For example, the French AI strategy proposes to study the labour market impacts and to implement ethics-by-design. Thus, the central topics of such societal dialogues include questions of the potential or real impact of AI and robotics on the work force and on society as a whole, and also questions of privacy, security, safety and adequate regulation.

12.1 The Role of Ethics

The question of what should be considered right and wrong in the development and deployment of AI and robotics is central to many published policy papers. For example, the European Commission (EC) now asks for the inclusion of *ethics in the development and use of new technologies* in programmes and courses and the development of ethical guidelines for the use and development of AI *in full respect of fundamental rights*. The EC plan goes as far as aiming to set a global ethical standard towards becoming a world leader in ethical, trusted AI. Whatever we may think of this aspiration, it is certainly true that there is a lack of practical, agreed guidelines and rules regarding systems that are much more autonomous in their calculations, actions and reactions than what we have been used to in the past. Such autonomous systems—or more specifically intelligent autonomous systems—act; they *do* things. Now, *what one should do* is the famous Kantian question underlying all ethical considerations.

Today, most jurisdictions around the world have only just started to investigate regulatory aspects of AI. In this book we have given preliminary answers to many of these issues, however industry has made the case that it requires clear rules for speedy innovation based on AI. Companies may steer away from AI applications in states of uncertainty in which the legal implications of bringing AI and robotic applications to the market are unclear.

Ethics therefore becomes important at many layers of the policy discussion. It is a topic for the engineer designing the system, including the student learning to build AI systems. It is also a topic for society to value the impacts of AI technology on the daily lives of citizens. Consequently, it is a key question for policy makers in discussions about AI and robotic technologies. Note that ethical aspects are not only discussed in questions of regulation. Much more importantly, ethical questions underpin the design of AI and robotic systems from defining the application to the details of their implementation. Ethics in AI is therefore much broader and concerns very basic design choices and considerations about which society we would like to live in.

12.2 International Cooperation

In its AI Action Plan, the EC identifies a need for coordinated action in ethics and in addressing societal challenges, but also the regulatory framework. It calls upon its member states to create synergies and cooperation on ethics. At the time of writing this book, countries around the world are looking for best practices in regulating—or deregulating—AI and robotics. Information and communication technologies in general have a tendency to generate impact across country borders. Already today, AI systems such as Google's translation service, for example, provide good quality translations to people all over the world. These translations are based on documents in many different languages available throughout the internet. In this way, Google exploits data that users publish only to create and improve its services often without people being aware that they support the development of improved translation services.

Many AI systems rely on massive amounts of data. And in many cases, this data may be considered personal. Questions of personal data protection have definitely become international ever since the EU defined its General Data Protection regulation (GDPR) to apply internationally, if not before. Privacy, data exchange, and AI are intimately related aspects that need to be put into an international context. There is a real international need to exchange concepts and ideas about how to best regulate or support AI and robotics. Countries will often take up regulatory models from other countries as they consider useful for their respective jurisdiction and the area of AI is no exception. Europe's GDPR has influenced policy makers world-wide, for example in the state of California. On the other hand, some countries may also decidedly vote against its underlying rationale and seek different legal approaches to data protection and privacy.

Similarly, the impact of AI on labour laws includes important international aspects. It is therefore not a coincidence that the New Zealand AI strategy, which remarkably is an industry association's paper, calls for more engagement with the international labour policy community. The New Zealand AI strategy also includes the challenging topic of AI for warfare. It is evident that this is a specific topic that needs to be discussed and perhaps regulated internationally.

Finally, international aspects do not stop with questions of regulation, labour or how to best design an AI system. Both the German and the EU strategies for AI include the important question which role AI should play in development policies. It is an yet open issue how to make sure that AI does not become a technology exclusively developed by a small set of industrialised leaders and instead also benefits developing countries.

References

Abdi, Jordan, Ahmed Al-Hindawi, Tiffany Ng, and Marcela P. Vizcaychipi. 2018. Scoping review on the use of socially assistive robot technology in elderly care. *British Medical Journal Open* 8: e018815. 10.1136/bmjopen-2017-018815. https://bmjopen.bmj.com/content/8/2/e018815. ISSN 2044-6055.

Adams, Thomas K. 2001. Future warfare and the decline of human decisionmaking. *Parameters* 31 (4): 57–71. https://ssi.armywarcollege.edu/pubs/parameters/articles/01winter/adams.pdf.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. 2019. Economic policy for artificial intelligence. *Innovation Policy and the Economy* 19: 139–159. https://doi.org/10.1086/699935.

Aldiss, Brian Wilson. 2001. *Supertoys last all summer long: and other stories of future time*. St. Martin's Griffin. http://www.worldcat.org/oclc/956323493. ISBN 978-0312280611.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *Pro Publica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Anton, A.I., J.B. Earp, J.D. Young, and 2010. How internet users' privacy concerns have evolved since, 2002. IEEE Security Privacy 8 (1): 21–27. *ISSN* 1540–7993: https://doi.org/10.1109/MSP.2010.38.

Arkin, Ronald C. 2008. Governing lethal behavior: embedding ethics in a hybrid deliberative/reactive robot architecture. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, 121–128. ACM. https://doi.org/10.1145/1349822.1349839. ISBN 978-1-60558-017-3.

Arkin, Ronald. 2009. *Governing lethal behavior in autonomous robots*. Chapman and Hall/CRC. http://www.worldcat.org/oclc/933597288. ISBN 978-1420085945.

Arkin, Ronald C. 2010. The case for ethical autonomy in unmanned systems. *Journal of Military Ethics* 9 (4): 332–341. https://doi.org/10.1080/15027570.2010.536402.

Arkin, Ronald Craig, Patrick Ulam, and Alan R. Wagner. 2012. Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception. *Proceedings of the IEEE* 100 (3): 571–589. https://doi.org/10.1109/JPROC.2011.2173265.

Arntz, Melanie, Terry Gregory, and Ulrich Zierahn. 2016. The risk of automation for jobs in oecd countries. *OECD Social, Employment and Migration Working Papers*, 189. https://doi.org/10.1787/5jlz9h56dvq7-en.

Baase, Sare, and Timothy M. Henry. 2017. *A gift of fire social, legal, and ethical issues for computing technology*, 5th ed. Prentice Hall PTR. http://www.worldcat.org/oclc/1050275090. ISBN 9780134615271.

C. Bartneck et al., *An Introduction to Ethics in Robotics and AI*,
SpringerBriefs in Ethics, https://doi.org/10.1007/978-3-030-51110-4

Bartlett, Jamie. 2018. *The People vs Tech*. Penguin Random House. http://www.worldcat.org/oclc/1077483710. ISBN 978-1785039065.

Bekey, George A. 2005. *Autonomous robots: from biological inspiration to implementation and control*. MIT press. http://www.worldcat.org/oclc/800006294. ISBN 978-0262025782.

Bentham, Jeremy. 1996. *The collected works of Jeremy Bentham: An introduction to the principles of morals and legislation*. Clarendon Press. http://www.worldcat.org/oclc/909020336. ISBN 978-0198205166.

Blackburn, Simon. 2002. *Being good: A short introduction to ethics*. OUP Oxford. http://www.worldcat.org/oclc/945382272. ISBN 978-0192853776.

Bonabeau, Eric. 2007. Understanding and managing complexity risk. *MIT Sloan Management Review* 48 (4): 62. https://sloanreview.mit.edu/article/understanding-and-managing-complexity-risk/.

Bonnefon, Jean-François, Azim Shariff, and Iyad Rahwan. (2016). The social dilemma of autonomous vehicles. Science 352 (6293): 1573–1576. doi: https://doi.org/10.1126/science.aaf2654.ISSN 0036-8075.

Bourget, David, and David J. Chalmers. 2014. What do philosophers believe? Philosophical Studies 170 (3): 465–500. *ISSN* 1573–0883: https://doi.org/10.1007/s11098-013-0259-7.

Brandstetter, Jürgen, and C. Bartneck. 2017. Robots will dominate the use of our language. *Adaptive Behaviour* 25 (6): 275–288. https://doi.org/10.1177/1059712317731606.

Brandstetter, Jurgen, Eduardo B. Sandoval, Clay Beckner, and Christoph Bartneck. 2017. Persistent lexical entrainment in hri. In *ACM/IEEE international conference on human-robot interaction*, 63–72. ACM. https://doi.org/10.1145/2909824.3020257. ISBN 978-1-4503-4336-7.

Braun, Megan, and Daniel R. Brunstetter. 2013. Rethinking the criterion for assessing cia-targeted killings: Drones, proportionality and jus ad vim. *Journal of Military Ethics* 12 (4): 304–324. https://doi.org/10.1080/15027570.2013.869390.

Broadbent, Elizabeth. 2017. Interactions with robots: The truths we reveal about ourselves. *Annual Review of Psychology* 68: 627–652. https://doi.org/10.1146/annurev-psych-010416-043958.

Brooks, Rodney. 2017. Unexpected consequences of self driving cars. *Rodney Brooks Blog*. https://rodneybrooks.com/unexpected-consequences-of-self-driving-cars/.

Buolamwini, Joy, Inioluwa Deborah Raji. 2019. Actionable auditing: Investigating the impact of publically naming biased performance results of commercial ai product. In *Proceedings of the AAAI/ACM Conference On Artifical Intelligence, Ethics, And Society*. AIES. http://www.aies-conference.com/wp-content/uploads/2019/01/AIES-19_paper_223.pdf.

Calo, Christopher J., Nicholas Hunt-Bull, Lundy Lewis, and Ted Metzler. 2011. Ethical implications of using the paro robot. In *2011 AAAI Workshop (WS-2011-2012)*, 20–24. AAAI. http://dl.acm.org/citation.cfm?id=2908724.2908728.

Carpenter, Julie. 2016. *Culture and human-robot interaction in militarized spaces: A war story*. London: Routledge.

Carsten, Oliver, and Marieke H. Martens. 2018. How can humans understand their automated cars? hmi principles, problems and solutions. Cognition, Technology & Work. doi: https://doi.org/10.1007/s10111-018-0484-0.ISSN 1435-5566.

Chouldechova, Alexandra. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments.

Cicero, Marcus Tullius. *De officiis* 44BC. https://www.gutenberg.org/ebooks/47001.

Coeckelbergh, Mark. 2009. Personal robots, appearance, and human good: a methodological reflection on roboethics. *International Journal of Social Robotics* 1 (3): 217–221. https://doi.org/10.1007/s12369-009-0026-2.

Coeckelbergh, Mark. 2010. Robot rights? towards a social-relational justification of moral consideration. Ethics and Information Technology 12 (3): 209–221. doi: https://doi.org/10.1007/s10676-010-9235-5..

Contag, M., G. Li, A. Pawlowski, F. Domke, K. Levchenko, T. Holz, and S. Savage. 2017. How they did it: An analysis of emission defeat devices in modern automobiles. In *2017 IEEE Symposium on Security and Privacy (SP)*, 231–250, May. https://doi.org/10.1109/SP.2017.66. ISBN 978-1-5090-5533-3.

Crane, A., and D. Matten. 2007. *Business ethics. Managing corporate citizenship and sustainability in the age of globalization*. Oxford University Press. http://www.worldcat.org/oclc/982687792. ISBN 978-0199697311.

Crew, Bec. 2015. Driverless cars could reduce traffic fatalities by up to 90 report. *Science Alert*. https://www.sciencealert.com/driverless-cars-could-reduce-traffic-fatalities-by-up-to-90-says-report.

Cummings, Missy. 2017. The brave new world of driverless cars. *TR News* 308: 34–37. https://trid.trb.org/view/1467060.

Darwall, Stephen. 1997. *Philosophical ethics: An historical and contemporary introduction*. London: Routledge. http://www.worldcat.org/oclc/1082497213. ISBN 978-0813378602.

Delvaux, Mady. 2017. *Report with recommendations to the commission on civil law rules on robotics*. Technical report A8-0005/2017, European Parliament. http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+REPORT+A8-2017-0005+0+DOC+XML+V0//EN.

Department of Defense. 2012. *Sustaining US global leadership: Priorities for the 21st century*. Department of Defense. http://archive.defense.gov/news/Defense_Strategic_Guidance.pdf. ISBN 978-1502887320.

DePaulo, Bella M., Deborah A. Kashy, Susan E. Kirkendol, Melissa M. Wyer, and Jennifer A. Epstein. 1996. Lying in everyday life. *Journal of Personality and Social Psychology* 70 (5): 979–995. https://doi.org/10.1037/0022-3514.70.5.979.

Desmon-Jhu, Stephanie. 2016. More college students are using adderall to stay up and study. *Futurity*. https://www.futurity.org/college-students-adderall-1107612-2/.

Diehl, Joshua J., Lauren M. Schmitt, Michael Villano, and Charles R. Crowell. 2012. The clinical use of robots for individuals with autism spectrum disorders: A critical review. *Research in Autism Spectrum Disorders* 6 (1): 249–262. https://doi.org/10.1016/j.rasd.2011.05.006.

Dieterich, William, Christina Mendoza, and Tim. Brennan. 2016. Compas risk scales: Demonstrating accuracy equity and predictive parity. *Equivant*. https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.

Dressel, Julia, and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4 (1). https://doi.org/10.1126/sciadv.aao5580. https://advances.sciencemag.org/content/4/1/eaao5580.

Duff, Robin Antony. 2007. *Answering for crime: Responsibility and liability in the criminal law*. Hart Publishing. ISBN 978-1849460330. http://www.worldcat.org/oclc/1073389374.

Ehrenkranz, Melanie. 2018. Houston votes to ban businesses from letting people screw humanlike devices in-store. *Gizomodo*. https://www.gizmodo.com.au/2018/10/houston-votes-to-ban-businesses-from-letting-people-screw-humanlike-devices-in-store/.

Enemark, Christian. 2013. *Armed drones and the ethics of war: military virtue in a post-heroic age*. London: Routledge. http://www.worldcat.org/oclc/896601866. ISBN 978-1138900882.

Ernest, Nicholas, C. David Carroll, M. Schumacher, K.Cohen Clark, and G. Lee. 2016. Genetic fuzzy based artificial intelligence for unmanned combat aerial vehicle control in simulated air combat missions. *Journal of Defense Management* 6 (144): 2167–0374. https://doi.org/10.4172/2167-0374.1000144.

Fantz, Ashley. 2015. Prison time for some atlanta school educators in cheating scandal. *CNN*. https://www.wired.com/story/self-driving-cars-rand-report/.

Federal Ministry of Transportation and Digital Infrastructure. 2017. Ethics commission - automated and connected driving. *BMVI*. https://www.bmvi.de/SharedDocs/EN/Documents/G/ethic-commission-report.pdf?__blob=publicationFile.

Feldman, Robert S., James A. Forrest, and Benjamin R. Happ. 2002. Self-presentation and verbal deception: Do self-presenters lie more? *Basic and Applied Social Psychology* 24 (2): 163–170. https://doi.org/10.1207/S15324834BASP2402_8.

Fischer, Manfred M., and Josef Fröhlich. 2013. *Knowledge, complexity and innovation systems*. Springer Science & Business Media. http://www.worldcat.org/oclc/906244357. ISBN 978-3540419693.

Floridi, Luciano. 2008. Artificial intelligence's new frontier: Artificial companions and the fourth revolution. *Metaphilosophy* 39 (4–5): 651–655. https://doi.org/10.1111/j.1467-9973.2008.00573.x.

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. 2018. Ai4people–an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. Minds and Machines 28 (4): 689–707. doi: https://doi.org/10.1007/s11023-018-9482-5.ISSN 1572-8641.

Fong, Terrence, Illah Nourbakhsh, and Kerstin Dautenhahn. 2003. A survey of socially interactive robots. *Robotics and Autonomous Systems* 42 (3–4): 143–166. https://doi.org/10.1016/S0921-8890(02)00372-X.

Ford, Martin. 2015. *Rise of the robots: technology and the threat of a jobless future*. Basic Books. http://www.worldcat.org/oclc/920676465. ISBN 978-0465097531.

Frey, Carl Benedikt, and Michael A. Osborne. 2017. The future of employment. *Technological Forecasting and Social Change* 114: 254–280. https://doi.org/10.1016/j.techfore.2016.08.019.

Goldman, Robert, Ronald Klatz, and Patricia J. Bush. 1987. *Death in the locker room*. Elite Sports Medicine Publications. http://www.worldcat.org/oclc/762212483. ISBN 978-0895865977.

Gonzalez, Robbie. 2017. Virtual therapists help veterans open up about ptsd. *Wired*. https://www.wired.com/story/virtual-therapists-help-veterans-open-up-about-ptsd/.

Goodall, Noah J. 2016. Can you program ethics into a self-driving car? IEEE Spectrum 53 (6): 28–58. *ISSN* 0018–9235: https://doi.org/10.1109/MSPEC.2016.7473149.

Goodrich, Michael A., Alan C. Schultz, et al. 2008. Human-robot interaction: A survey. *Foundations and Trends in Human-Computer Interaction* 1 (3): 203–275. https://doi.org/10.1561/1100000005.

Gorr, Wilpen L., Daniel Nagin, and Janusz Szczypula. 1994. Comparative study of artificial neural network and statistical models for predicting student grade point averages. International Journal of Forecasting 10 (1): 17–34. doi: https://doi.org/10.1016/0169-2070(94)90046-9.ISSN 0169-2070.

Greenberg, Andy. 2015. Hackers remotely kill a jeep on the highway—with me in it. *Wired*. https://www.wired.com/2015/07/hackers-remotely-kill-jeep-highway/.

Gunkel, David J. 2018. *Robot Rights*. MIT Press. http://www.worldcat.org/oclc/1047850795. ISBN 978-0262038621.

Güth, Werner, Rolf Schmittberger, and Bernd Schwarze. 1982. An experimental analysis of ultimatum bargaining. Journal of Economic Behavior & Organization 3 (4): 367–388. doi: https://doi.org/10.1016/0167-2681(82)90011-7.ISSN 0167-2681.

Hancock, Peter A., Deborah R. Billings, Kristin E. Schaefer, Jessie Y.C. Chen, Ewart J. De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors* 53 (5): 517–527. https://doi.org/10.1177/0018720811417254.

Haring, Kerstin Sophie, Céline Mougenot, Fuminori Ono, and Katsumi Watanabe. 2014a. Cultural differences in perception and attitude towards robots. *International Journal of Affective Engineering* 13 (3): 149–157. https://doi.org/10.5057/ijae.13.149.

Haring, Kerstin Sophie, David Silvera-Tawil, Yoshio Matsumoto, Mari Velonaki, and Katsumi Watanabe. 2014b. Perception of an android robot in Japan and Australia: A cross-cultural comparison. In *International Conference on Social Robotics*, 166–175. Springer. https://doi.org/10.1007/978-3-319-11973-1_17. ISBN 978-3-319-11972-4.

Hawkins, Andrew J. 2018. Elon musk still doesn't think lidar is necessary for fully driverless cars. *The Verge*. https://www.theverge.com/2018/2/7/16988628/elon-musk-lidar-self-driving-car-tesla.

Heider, Fritz, and Marianne Simmel. 1944. An experimental study of apparent behavior. *The American Journal of Psychology* 57 (2): 243–259. https://doi.org/10.2307/1416950.

Herman, Arthur. 2018. China's brave new world of AI. *Forbes*. https://www.forbes.com/sites/arthurherman/2018/08/30/chinas-brave-new-world-of-ai/.

Heyns, Christof. 2016. Autonomous weapons systems: Living a dignified life and dying a dignified death, 3–20. *Cambridge University Press*. https://doi.org/10.1017/CBO9781316597873.001.

Hildrenbrand, Jerry. 2018. Amazon alexa: What kind of data does amazon get from me? *Android Central*. https://www.androidcentral.com/amazon-alexa-what-kind-data-does-amazon-get-me.

Holpuch, Amanda. 2016. Tim cook says apple's refusal to unlock iphone for FBI is a 'civil liberties' issue. *The Guardian*. https://www.theguardian.com/technology/2016/feb/22/tim-cook-apple-refusal-unlock-iphone-fbi-civil-liberties.

Hu, Liming, David Bell, Sameer Antani, Zhiyun Xue, Kai Yu, Matthew P Horning, Noni Gachuhi, Benjamin Wilson, Mayoore S Jaiswal, Brian Befano, L Rodney Long, Rolando Herrero, Mark H Einstein, Robert D Burk, Maria Demarco, Julia C Gage, Ana Cecilia Rodriguez, Nicolas Wentzensen, and Mark Schiffman. 2019. *An observational study of deep learning and automated evaluation of cervical images for cancer screening*. http://oup.prod.sis.lan/jnci/advance-article-pdf/doi/10.1093/jnci/djy225/27375757/djy225.pdf. https://doi.org/10.1093/jnci/djy225.

Ingram, Brandon, Daniel Jones, Andrew Lewis, Matthew Richards, Charles Rich, and Lance Schachterle. 2010. A code of ethics for robotics engineers. In *Proceedings of the 5th ACM/IEEE international conference on human-robot interaction*, 103–104. IEEE Press. https://doi.org/10.1109/HRI.2010.5453245. ISBN 978-1-4244-4892-0.

International Committee of the Red Cross. 2015. What are jus ad bellum and jus in bello. *International Committee of the Red Cross*. https://www.icrc.org/en/document/what-are-jus-ad-bellum-and-jus-bello-0.

James, Vincent. 2017. Putin says the nation that leads in ai 'will be the ruler of the world'. *The Verge*. https://www.theverge.com/2017/9/4/16251226/russia-ai-putin-rule-the-world.

Jr, Kahn, H. Peter, Takayuki Kanda, Hiroshi Ishiguro, Brian T. Gill, Jolina H. Ruckert, Solace Shen, Heather E. Gary, Aimee L. Reichert, Nathan G. Freier, and Rachel L. Severson. 2012. *Do people hold a humanoid robot morally accountable for the harm it causes? Proceedings of the seventh annual ACM/IEEE international conference on human-robot interaction, HRI '12, 33–40*. New York: ACM. https://doi.org/10.1145/2157689.2157696. ISBN 978-1-4503-1063-5.

Kalra, Nidhi, and Susan M. Paddock. 2016. Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? *Transportation Research Part A: Policy and Practice* 94: 182–193. https://doi.org/10.1016/j.tra.2016.09.010.

Kant, Immanuel. 1785. *Groundwork of the metaphysic of morals*. http://www.worldcat.org/oclc/1057708209.

Kant, Immanuel. 1788. *The critique of practical reason*. https://www.gutenberg.org/ebooks/5683.

Kaplan, Andreas, and Michael Haenlein. 2019. Siri, siri, in my hand: Who's the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence. Business Horizons 62 (1): 15–25. doi: https://doi.org/10.1016/j.bushor.2018.08.004.ISSN 0007-6813.

Kaste, Martin. 2018. Facebook increasingly reliant on a.i. to predict suicide risk. *NPR*. https://www.npr.org/2018/11/17/668408122/facebook-increasingly-reliant-on-a-i-to-predict-suicide-risk.

Kidd, Cory D., and Cynthia Breazeal. 1999, 2007. A robotic weight loss coach. In *Proceedings of the national conference on artificial intelligence*, vol. 22, 1985. Menlo Park, CA; Cambridge, MA; London: AAAI Press; MIT Press; AAAI. https://www.aaai.org/Papers/AAAI/2007/AAAI07-366.pdf.

Kim, Ki Joon, Eunil Park, and S. Shyam Sundar. 2013. Caregiving role in human-robot interaction: A study of the mediating effects of perceived benefit and social presence. *Computers in Human Behavior* 29 (4): 1799–1806. http://www.sciencedirect.com/science/article/pii/S0747563213000757. https://doi.org/10.1016/j.chb.2013.02.009. ISSN 0747-5632.

Klein, Martha. 2005. Responsibility. In *The Oxford companion to philosophy*, ed. Ted Honderich. OUP Oxford. http://www.worldcat.org/oclc/180031201. ISBN 978-0199264797.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. Algorithmic fairness. *Aea papers and proceedings* 108: 22–27.

Kleinz, Torsten. 2018. Versteigerte werbeplätze: So funktioniert programmatic advertising. *Heise*. https://www.heise.de/ct/artikel/Versteigerte-Werbeplaetze-So-funktioniert-Programmatic-Advertising-4203227.html.

Kolb, Michael. 2012. *Soldier and robot interaction in combat environments*. Ph.D. thesis, The University of Oklahoma. https://ou-primo.hosted.exlibrisgroup.com/primo-explore/fulldisplay?docid=NORMANLAW_ALMA21340006270002042&context=L&vid=OUNEW&lang=en_US.

Kolowich, Steve. 2012. Recommended for you. *Inside Higher Education*. https://www.insidehighered.com/news/2012/03/16/university-builds-course-recommendation-engine-steer-students-toward-completion.

Laris, Michael. 2018. A waymo safety driver collided with a motorcyclist. The company says a self-driving minivan would have done better. *Washington Post*. https://www.washingtonpost.com/technology/2018/11/06/waymo-safety-driver-collides-with-motorcyclist-company-says-self-driving-minivan-would-have-done-better/?utm_term=.e48719a57a13.

Lee, John D., and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors* 46 (1): 50–80. https://doi.org/10.1518/hfes.46.1.50_30392. PMID: 15151155.

Lee, Sang M., and Silvana Trimi, 2018. Innovation for creating a smart future. Journal of Innovation & Knowledge 3 (1): 1–8. doi: https://doi.org/10.1016/j.jik.2016.11.001.ISSN 2444-569X.

Li, Bo-hu, Bao-cun Hou, Wen-tao Yu, Xiao-bing Lu, and Chun-wei Yang. 2017. Applications of artificial intelligence in intelligent manufacturing: a review. Frontiers of Information Technology & Electronic Engineering 18 (1): 86–96. doi: https://doi.org/10.1631/FITEE.1601885.ISSN 2095-9230.

Lin, Patrick. 2016. Why ethics matters for autonomous cars. In *Autonomous driving*, ed. Markus Maurer, J. Christian Gerdes, Barbara Lenz, Hermann Winner, et al., 70–85. Berlin: Springer. https://doi.org/10.1007/978-3-662-48847-8_4. ISBN 978-3-662-48845-4.

Lin, Patrick, Keith Abney, and Ryan Jenkins. 2017. *Robot ethics 2.0: From autonomous cars to artificial intelligence*. Oxford University Press. http://www.worldcat.org/oclc/1011372036. ISBN 978-0190652951.

Lin, Zhiyuan, Alex Chohlas-Wood, and Sharad Goel. 2019. Guiding prosecutorial decisions with an interpretable statistical model. In *Proceedings of the AAAI/ACM conference on artifical intelligence, ethics, and society*. https://footprints.stanford.edu/papers/smart-prosecution.pdf.

Lins, Karl V., Henri Servaes, and Ane Tamayo. 2017. Social capital, trust, and firm performance: The value of corporate social responsibility during the financial crisis. *The Journal of Finance* 72 (4): 1785–1824. https://doi.org/10.1111/jofi.12505.

Lokhorst, Gert-Jan, and Jeroen Van Den Hoven. 2012. Responsibility for military robots (Chap. 9). In *Robot ethics: The ethical and social implications of robotics*, ed. Patrick Lin and George A Bekey, 145–156. Cambridge: MIT Press. http://www.worldcat.org/oclc/978497330. ISBN 978-0262526005.

Luetge, Christoph. 2013. Handbook of the philosophical foundations of business ethics. https://doi.org/10.1007/978-94-007-1494-6.

Luetge, Christoph. 2017. The german ethics code for automated and connected driving. *Philosophy & Technology* 30 (4): 547–558. https://doi.org/10.1007/s13347-017-0284-0. ISSN 2210-5441.

Luetge, Christoph, Eberhard Schnebel, and Nadine Westphal. 2014. Risk management and business ethics: Integrating the human factor. In *Risk: A multidisciplinary introduction*, ed. Claudia Klüppelberg, Daniel Straub, and Isabell Welpe, 37–61. Springer. https://doi.org/10.1007/978-3-319-04486-6.

Luetke, Christoph. 2017. Responsibilities of online service providers from a business ethics point of view. In *The responsibilities of online service providers*, ed. Mariarosaria Taddeo and Luciano Floridi, 119–133. Springer. https://doi.org/10.1007/978-3-319-47852-4_7.

Lütge, Christoph, Hannes Rusch, and Matthias Uhl, and Christoph Luetge. 2014. *Experimental ethics: Toward an empirical moral philosophy*. Palgrave Macmillan. http://www.worldcat.org/oclc/896794689. ISBN 978-1349488797.

Mackie, John. 1991. *Ethics: Inventing right and wrong*. UK: Penguin. http://www.worldcat.org/oclc/846989284. ISBN 978-0140135589.

Makridakis, Spyros. 2017. The forthcoming artificial intelligence (ai) revolution: Its impact on society and firms. Futures 90: 46–60. doi: https://doi.org/10.1016/j.futures.2017.03.006.ISSN 0016-3287.

Malle, Bertram F., Matthias Scheutz, and Joseph L. Austerweil. 2017. Networks of social and moral norms in human and robot agents. In *A world with robots*, 3–17. Springer.

Marshall, Aarian. 2017. To save the most lives, deploy (imperfect) self-driving cars asap. *Wired*. https://www.wired.com/story/self-driving-cars-rand-report/.

Marshall, Aarian. 2018. We've been talking about self-driving car safety all wrong. *Wired*. https://www.wired.com/story/self-driving-cars-safety-metrics-miles-disengagements/.

Martinez-Conesa, Isabel, Pedro Soto-Acosta, and Mercedes Palacios-Manzano. 2017. Corporate social responsibility and its effect on innovation and firm performance: An empirical research in smes. *Journal of Cleaner Production* 142 (4): 2374–2383. https://doi.org/10.1016/j.jclepro.2016.11.038.

Michalski, Ryszard S., Jaime G. Carbonell, and Tom M. Mitchell. 2013. *Machine learning: An artificial intelligence approach*. Springer Science & Business Media. URL http://www.worldcat.org/oclc/864590508. ISBN 978-3662124079.

Mill, John Stuart. 1863. *Utilitarianism*. London: Parker, Son and Bourn. https://www.gutenberg.org/ebooks/11224.

Moor, J. H. 2006. The nature, importance, and difficulty of machine ethics. IEEE Intelligent Systems 21 (4): 18–21. doi: https://doi.org/10.1109/MIS.2006.80.ISSN 1541-1672.

Moseley, Laurence G., and Donna M. Mead. 2008. Predicting who will drop out of nursing courses: A machine learning exercise. Nurse Education Today 28 (4): 469–475. doi: https://doi.org/10.1016/j.nedt.2007.07.012.ISSN 0260-6917.

National Police Foundation. 2017. A review of the baltimore police department's use of persistent surveillance. *National Police Foundation*. https://www.policefoundation.org/publication/a-review-of-the-baltimore-police-departments-use-of-persistent-surveillance/.

Nourbakhsh, Illah Reza. 2013. *Robot futures*. Cambridge: MIT Press. http://www.worldcat.org/oclc/1061811057. ISBN 978-0262528320.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366 (6464): 447–453.

Ogawa, Kohei, Christoph Bartneck, Daisuke Sakamoto, Takayuki Kanda, T. Ono, and Hiroshi Ishiguro. 2009. Can an android persuade you? In 18th IEEE international symposium on robot and human interactive communication, RO-MAN2009, 553–557. IEEE. doi: https://doi.org/10.1109/ROMAN.2009.5326352.ISBN 978-1-4244-5081-7.

O'Kane, Sean. 2018. Uber reportedly thinks its self-driving car killed someone because it 'decided' not to swerve. *The Verge*. https://www.theverge.com/2018/5/7/17327682/uber-self-driving-car-decision-kill-swerve.

Perkowitz, Sidney. 2004. *Digital people: From bionic humans to androids*. Joseph Henry Press. http://www.worldcat.org/oclc/936950712. ISBN 978-0309096195.

Pies, Ingo. 2010. Sustainability in the petroleum industry: Theory and practice of voluntary self-commitments. *University of Wittenberg Business Ethics Study, No*. 2010–1: https://doi.org/10.2139/ssrn.1595943.

Poole, David L., and Alan K. Mackworth. 2010. Artificial intelligence: Foundations of computational agents. Cambridge University Press. doi: https://doi.org/10.1017/CBO9780511794797.ISBN 9780511794797.

Povyakalo, Andrey A., Eugenio Alberdi, Lorenzo Strigini, and Peter Ayton. 2013. How to discriminate between computer-aided and computer-hindered decisions: a case study in mammography. *Medical Decision Making* 33 (1): 98–107. https://doi.org/10.1177/0272989X12465490.

Quinn, Michael J. 2017. *Ethics for the information age*, 7th ed. Addison-Wesley Publishing Company. http://www.worldcat.org/oclc/1014043739. ISBN 978-0134296548.

Reeves, Byron, and Clifford Ivar Nass. 1996. *The media equation: how people treat computers, televisions, and new media like real people and places*. Stanford, Calif. New York; Cambridge: CSLI Publications; Cambridge University Press. http://www.worldcat.org/oclc/1061025314. ISBN 978-1575860534.

Reynolds, Emily. 2018. The agony of sophia, the world's first robot citizen condemned to a lifeless career in marketing. *Wired*. https://www.wired.co.uk/article/sophia-robot-citizen-womens-rights-detriot-become-human-hanson-robotics.

Richardson, Kathleen. 2016. The asymmetrical'relationship': parallels between prostitution and the development of sex robots. *ACM SIGCAS Computers and Society* 45 (3): 290–293. https://doi.org/10.1145/2874239.2874281.

Riek, Laurel, and Don Howard. 2014. A code of ethics for the human-robot interaction profession. *Proceedings of We robot*. https://ssrn.com/abstract=2757805.

Robinette, Paul, Wenchen Li, Robert Allen, Ayanna M. Howard, and Alan R. Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *Proceedings of the eleventh ACM/IEEE international conference on human robot interaction*, 101–108. IEEE Press. https://doi.org/10.1109/HRI.2016.7451740. ISBN 978-1-4673-8370-7.

Rothenbücher, Dirk, Jamy Li, David Sirkin, Brian Mok, and Wendy Ju. 2016. Ghost driver: A field study investigating the interaction between pedestrians and driverless vehicles. In *25th IEEE international symposium on robot and human interactive communication (RO-MAN)*, 795–802. IEEE. https://doi.org/10.1109/ROMAN.2016.7745210. ISBN 978-1-5090-3929-6.

Russell, Stuart J., and Peter Norvig. 2010. *Artificial intelligence: a modern approach*, 3rd ed. Upper Saddle River, N.J.: Prentice Hall. http://www.worldcat.org/oclc/688385283. ISBN 9780132071482.

Saeidi, Sayedeh Parastoo, Saudah Sofian, Parvaneh Saeidi, Sayyedeh Parisa Saeidi, and Seyyed Alireza Saaeidi. 2015. How does corporate social responsibility contribute to firm financial performance? the mediating role of competitive advantage, reputation, and customer satisfaction. *Journal of business research* 68 (2): 341–350. https://doi.org/10.1016/j.jbusres.2014.06.024.

Scharre, Paul. 2018. *Army of none: Autonomous weapons and the future of war*. WW Norton & Company.

Scheutz, Matthias. 2014. *The inherent dangers of unidirectional emotional bonds between humans and social robots* (Chap. 13), 205–222. Cambridge: MIT Press. http://www.worldcat.org/oclc/978497330. ISBN 9780262526005.

Searle, John R. 1980. Minds, brains and programs. *Behavioral and Brain Sciences* 3 (3): 417–457. https://doi.org/10.1017/S0140525X00005756.

Seifert, Inessa, Matthias Bürger, Leo Wangler, Stephanie Christmann-Budian, Marieke Rohde, Peter Gabriel, and Guido Zinke. 2018. Potenziale der künstlichen intelligenz im produzierenden gewerbe. (potentials of ai in manufacturing). https://www.bmwi.de/Redaktion/DE/Publikationen/Studien/potenziale-kuenstlichen-intelligenz-im-produzierenden-gewerbe-in-deutschland.pdf?__blob=publicationFile&v=8.

Sharkey, Noel. 2010. Saying 'no!' to lethal autonomous targeting. *Journal of Military Ethics* 9 (4): 369–383. https://doi.org/10.1080/15027570.2010.537903.

Sharkey, Noel, and Amanda Sharkey. 2010. The crying shame of robot nannies: an ethical appraisal. *Interaction Studies* 11 (2): 161–190. https://doi.org/10.1075/is.11.2.01sha.

Sharkey, Amanda, and Noel Sharkey. 2012. Granny and the robots: ethical issues in robot care for the elderly. *Ethics and Information Technology* 14 (1): 27–40. https://doi.org/10.1007/s10676-010-9234-6.

Shields, Nicholas. 2018. New survey shows consumers are wary of smart home devices invading their privacy. *Business Insider*. https://www.businessinsider.com/survey-says-consumers-have-privacy-concerns-with-smart-home-devices-2018-4.

Shiomi, Masahiro, Kazuhiko Shinozawa, Yoshifumi Nakagawa, Takahiro Miyashita, Toshio Sakamoto, Toshimitsu Terakubo, Hiroshi Ishiguro, and Norihiro Hagita. 2013. Recommendation effects of a social robot for advertisement-use context in a shopping mall. *International Journal of Social Robotics* 5 (2): 251–262. https://doi.org/10.1007/s12369-013-0180-4.

Shiu, Yung-Ming, and Shou-Lin Yang. 2017. Does engagement in corporate social responsibility provide strategic insurance-like effects? *Strategic Management Journal* 38 (2): 455–470. https://onlinelibrary.wiley.com/doi/abs/10.1002/smj.2494.

Short, E., J. Hart, M. Vu, and B. Scassellati. 2010. No fair!! an interaction with a cheating robot. In *Proceedings of the 5th ACM/IEEE international conference on human-robot interaction (HRI)*, 219–226. https://doi.org/10.1109/HRI.2010.5453193.

Singer, Peter Warren. 2009. *Wired for war: The robotics revolution and conflict in the twenty-first century*. Penguin. http://www.worldcat.org/oclc/958145424. ISBN 1594201986.

Sloman, Aaron, and Monica Croucher. 1981. Why robots will have emotions. In *Proceedings of the 7th international joint conference on artificial intelligence*, vol. 1, 197–202. http://dl.acm.org/citation.cfm?id=1623156.1623194.

Sparrow, Robert. 2016. Kicking a robot dog. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*, 229–229. https://doi.org/10.1109/HRI.2016.7451756.

Sparrow, Robert. 2017. Robots, rape, and representation. International Journal of Social Robotics 9 (4): 465–477. doi: https://doi.org/10.1007/s12369-017-0413-z.ISSN 1875-4805.

Sparrow, Robert, and Mark Howard. 2017. When human beings are like drunk robots: Driverless vehicles, ethics, and the future of transport. *Transportation Research Part C: Emerging Technologies* 80: 206–215. https://doi.org/10.1016/j.trc.2017.04.014.

Stephens-Davidowitz, Seth, and Steven Pinker. 2017. *Everybody lies: big data, new data, and what the internet can tell Us about who we really are*. New York: HarperCollins. http://www.worldcat.org/oclc/1004000087. ISBN 978-0062390851.

Surden, Harry, and Mary-Anne Williams. 2016. Technological opacity, predictability, and self-driving cars. *Cardozo Law Review* 38: 1–52. http://hdl.handle.net/10453/120588.

Thielmann, Sam. 2016. Use of police robot to kill dallas shooting suspect believed to be first in us history. *The Guardian*. https://www.theguardian.com/technology/2016/jul/08/police-bomb-robot-explosive-killed-suspect-dallas.

Thompson, Dennis F. 1980. Moral responsibility of public officials: The problem of many hands. *The American Political Science Review* 74 (4): 905–916. https://doi.org/10.2307/1954312.

Toma, Catalina L., Jeffrey T. Hancock, and Nicole B. Ellison. 2008. Separating fact from fiction: An examination of deceptive self-presentation in online dating profiles. *Personality and Social Psychology Bulletin* 34 (8): 1023–1036. https://doi.org/10.1177/0146167208318067.

United Nations. 1968. *Convention on road traffic*. United Nations. https://treaties.un.org/doc/Publication/MTDSG/Volume%20I/Chapter%20XI/XI-B-19.en.pdf.

van Wynsberghe, Aimee, and Scott Robbins. 2019. Critiquing the reasons for making artificial moral agents. *Science and Engineering Ethics* 25 (3): 719–735.

Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. 2017. Transparent, explainable, and accountable AI for robotics. *Science Robotics* 2 (6): eaan6080. https://doi.org/10.1126/scirobotics.aan6080.

Wallach, Wendell. 2008. *and Colin Allen*. Moral machines: Teaching robots right from wrong. Oxford University Press.

Wang, Lin, Pei-Luen Patrick Rau, Vanessa Evers, Benjamin Krisper Robinson, and Pamela Hinds. 2010. When in rome: the role of culture & context in adherence to robot recommendations. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*, 359–366. IEEE Press. https://doi.org/10.1109/HRI.2010.5453165. ISBN 978-1-4244-4892-0.

Wang, Qian, Junsheng Dou, and Shenghua Jia. 2016. A meta-analytic review of corporate social responsibility and corporate financial performance: The moderating effect of contextual factors. *Business & Society* 55 (8): 1083–1121. https://doi.org/10.1177/0007650315584317.

Watanabe, Miki, Kohei Ogawa, and Hiroshi Ishiguro. 2015. Can androids be salespeople in the real world? In *Proceedings of the 33rd annual ACM conference extended abstracts on human factors in computing systems*, 781–788. ACM. https://doi.org/10.1145/2702613.2702967. ISBN 978-1-4503-3146-3.

Weinberger, David. 2018. Don't make AI artificially stupid in the name of transparency. *Wired*. https://www.wired.com/story/dont-make-ai-artificially-stupid-in-the-name-of-transparency/.

Welsh, Sean. 2018. *Ethics and security automata. Ethics, emerging technologies and international affairs*. Abingdon: Routledge. http://www.worldcat.org/oclc/1004169786. ISBN 978-1138050228.

White, Daniel K., Christine E. Peloquin, Tuhina Neogi, Yanyan Zhu, Yuqing Zhang, Michael C. Nevitt, Elsa S. Strotmeyer, Jane A. Cauley, Robert M. Boudreau, Eleanor M. Simonsick, Luigi Ferrucci, Tamara B. Harris, and Susan M. Satterfield. 2012. Trajectories of gait speed predict mortality in well-functioning older adults: The health, aging and body composition study. *The Journals of Gerontology: Series A* 68 (4): 456–464. https://doi.org/10.1093/gerona/gls197.

Wilens, Timothy E., Lenard A. Adler, Jill Adams, Stephanie Sgambati, John Rotrosen, Robert Sawtelle, Linsey Utzinger, and Steven Fusillo. 2008. Misuse and diversion of stimulants prescribed for ADHD: A systematic review of the literature. *Journal of the American Academy of Child & Adolescent Psychiatry* 47 (1): 21–31. https://doi.org/10.1097/chi.0b013e31815a56f1.

Woolley, S. 2005. Children of Jehovah's witnesses and adolescent Jehovah's witnesses: what are their rights? *Archives of Disease in Childhood* 90 (7): 715–719. 10.1136/adc.2004.067843. https://adc.bmj.com/content/90/7/715. ISSN 0003-9888.

Zalta, Edward N. 2003. *Stanford encyclopedia of philosophy*. https://plato.stanford.edu/.

Zlotowski, Jakub, Diane Proudfoot, Kumar Yogeeswaran, and Christoph Bartneck. 2015. Anthropomorphism: Opportunities and challenges in human-robot interaction. *International Journal of Social Robotics* 7 (3): 347–360. https://doi.org/10.1007/s12369-014-0267-6.

Index

C. Bartneck et al., *An Introduction to Ethics in Robotics and AI*,
SpringerBriefs in Ethics, https://doi.org/10.1007/978-3-030-51110-4