sienna.

D4.4: Ethical Analysis of AI and Robotics Technologies

WP4: AI & robotics – Ethical, legal and social analysis

Lead authors	Philip Jansen, *University of Twente* (p.h.jansen@utwente.nl) Philip Brey, *University of Twente* (p.a.e.brey@utwente.nl)
Other authors	Alice Fox, Jonne Maas, Bradley Hillas, Nils Wagner, Patrick Smith, Isaac Oluoch, Laura Lamers, Hero van Gein, *University of Twente*; Anaïs Resseguier, Rowena Rodrigues, David Wright, *Trilateral Research Ltd.*; David Douglas, *Commonwealth Scientific and Industrial Research Organisation*
Contributors (country studies)	Marcelo de Araujo, Rachel Herdy, Fabio Shecaira, *Federal University of Rio de Janeiro*; Yuanyuan Zhang, Qian Wang, *Dalian University of Technology*; Robert Gianni, *SciencesPo Paris*; Lisa Tambornino, Philipp Hövel, *European Network of Research Ethics Committees Office*; Maria Bottis, *Ionian University*; Konrad Siemaszko, *Helsinki Foundation for Human Rights*; Jantina de Vries, Olivia Matshabane, *University of Cape Town*; Javier Valls Prieto, *University of Granada*; Gry Houeland, Heidi Howard, *Uppsala University*; Adam Holland, Chris Bavitz, *Berkman Klein Center for Internet & Society*
Reviewers & commenters	Alice Fox, *Virginia Polytechnic Institute and State University*; Ana Beduschi, *University of Exeter*; Cansu Canca, *AI Ethics Lab*; Christine Aicardi, *King's College London*; David Douglas, *Commonwealth Scientific and Industrial Research Organisation*; Diane Whitehouse, *European Health Telematics Association*; Jessica Sorenson, *Aarhus University*; John-Stewart Gordon, *Vytautas Magnus University*; Mark Coeckelbergh, *University of Vienna*

Due date	August 31st, 2019
Delivery date	August 31st, 2019
Type	Report
Dissemination level	PU

Keywords	Ethical issues; artificial intelligence; AI; robotics; robots; ethics; emerging technologies

The SIENNA project - *Stakeholder-informed ethics for new technologies with high socio-economic and human rights impact* - has received funding under the European Union's H2020 research and innovation programme under grant agreement No 741716.

Abstract

This SIENNA deliverable offers a broad ethical analysis of artificial intelligence (AI) and robotics technologies. Its primary aims have been to comprehensively identify and analyse the present and potential future ethical issues in relation to: (1) the AI and robotics subfields, techniques, approaches and methods; (2) their physical technological products and procedures that are designed for practical applications; and (3) the particular uses and applications of these products and procedures. In conducting the ethical analysis, we strove to provide ample clarification, details about nuances, and contextualisation of the ethical issues that were identified, while avoiding the making of moral judgments and proposing of solutions to these issues.

A secondary aim of this report has been to convey the results of SIENNA's "country studies" of the national academic and popular media debate on the ethical issues in AI and robotics in twelve different EU and non-EU countries, highlighting the similarities and differences between these countries. While these country study results have only formed a minor contribution to the overall identification and analysis of the ethical issues in this report, they are expected to play a larger role in future SIENNA deliverables.

This deliverable also provides an overview of the history and state of the art of the academic debate on ethics of AI and robot ethics, and an overview of the current institutional support of these fields.

Document history

Version	Date	Description	Reason for change	Distribution
V0.9	August 1st, 2019	Final draft report for external review	-	August 1st, 2019
V1.0	August 31st, 2019	Final report for submission to the EC	Reviews and comments	August 31st, 2019
V1.1	June 17th, 2020	Final report update	Corrected acknowledgement p.23	June 17th, 2020

Information in this report that may influence other SIENNA tasks

Linked task	Points of relevance
Task 4.7	The proposal for an ethical framework for AI and robotics will follow-up on the current report, as the framework will be based on important issues identified and analysed in this report.
Task 5.4	The code of responsible conduct for researchers in the fields of AI and robotics will require consideration of the issues identified in this report.
Task 6.1	The report on adapting methods for ethical analysis of emerging technologies will require contemplation about the successes and challenges in the methodology used to write this report.

Task 6.3	The step-by-step guidance from ethical analysis to ethical codes and operational guidelines task will require reflection about the successes and challenges in writing this report.

Contents

Executive summary

This report has been written for the SIENNA project, a European Union (EU) funded project which is part of the Horizon 2020 research and innovation programme. SIENNA aims to develop ethical frameworks, operational guidelines for research ethics committees, codes of responsible conduct and policy recommendations for new technologies with high socio-economic and human rights impacts. It also aims to develop general methods for the ethical and legal assessment of emerging technologies, and for the implementation of ethical frameworks and the development of policy recommendations. SIENNA focuses in particular on an assessment of three technology areas: (1) artificial intelligence (AI) and robotics; (2) human enhancement; and (3) human genomics.

Objectives and structure of the report

As part of the SIENNA project, this report engages in an extensive ethical analysis of AI and robotics technologies, including their various manifestations and applications. It aims to identify and analyse ethical issues in AI and robotics, both present and potential future ethical issues, with a time horizon of twenty years. The aim of the report is not to make recommendations or present solutions, but only to identify and analyse ethical issues. As such, the report stands on its own: it is a timely report, unique in its breadth and scope, that charts the ethical issues that should be taken into account in the development, use and regulation of AI and robotics technologies. In the context of SIENNA, it is also intended to provide a basis for our next report, in which we aim to present an ethical framework for AI and robotics that contains recommendations and solutions for ethical issues.

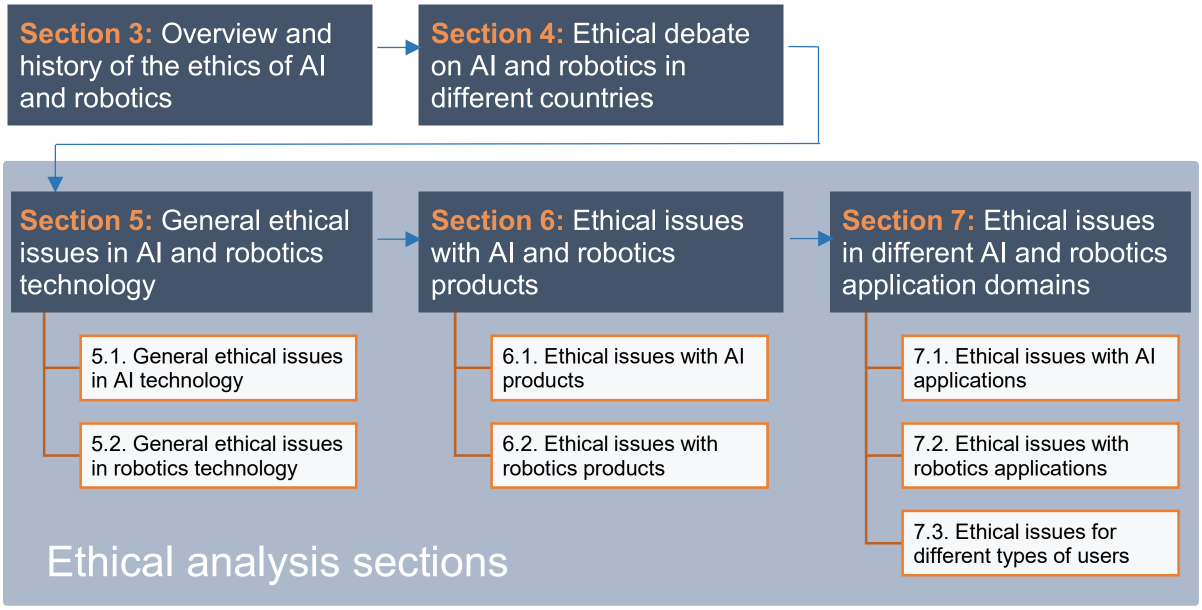

Section 3: Overview and history of the ethics of AI and robotics
Section 4: Ethical debate on AI and robotics in different countries
Section 5: General ethical issues in AI and robotics technology
5.1. General ethical issues in AI technology
5.2. General ethical issues in robotics technology
Section 6: Ethical issues with AI and robotics products
6.1. Ethical issues with AI products
6.2. Ethical issues with robotics products
Section 7: Ethical issues in different AI and robotics application domains
7.1. Ethical issues with AI applications
7.2. Ethical issues with robotics applications
7.3. Ethical issues for different types of users
Ethical analysis sections

Figure 1: Structure of the five substantive sections (3–7) of this report.

The report consists of five substantive parts (sections 3 through 7), next to an introduction (section 1), conclusion (section 8), and a section on methodology (section 2). Section 3 provides context to the ethical analysis that is to come, by providing a brief history of ethics of AI and robotics, covering both academic research and practical action. Also providing context, section 4 reviews how ethical issues in AI and robotics have been debated in different countries, both in the EU and globally. Sections 5 through 7 contain the actual ethical analysis, in three parts. Section 5 (part 1 of the ethical analysis) contains an analysis of general ethical issues in AI and robotics: issues that pertain to the technology in general, across its various manifestations and applications. Section 6 (part 2) considers ethical issues that apply to specific AI and robotics products and systems, such as intelligent agents, decision-support systems, social robots and drones. Section 7 (part 3), finally, considers ethical issues in particular application domains of AI and robotics, such as healthcare, education, law enforcement and defence. Figure 1 on the previous page provides an overview of the structure of the five substantive sections of this report.

In what follows, we first briefly present the methodology of our study, and then summarize the main results of the five substantive sections of our report (sections 3 through 7).

Methodology

The methodology for the ethical analysis of AI and robotics, carried out in sections 5, 6 and 7 of the report was developed earlier in the SIENNA project, and is called the "SIENNA approach to ethical analysis". It is based on literature review, consultation of experts and stakeholders, and original ethical analysis. It consists of a six-step process that is visualized in figure 2 at the beginning of section 2. In the report, five of these steps are carried out. The sixth step, recommendations and options for ethical decision-making, will be carried out in a later report.

In the first step, we specified the subject, aim and scope of analysis. During this step, we identified and defined the technologies, technological products, and application domains that we wanted to study, i.e., AI and robotics technologies, and their various manifestations and applications, both at present, and as they may evolve over the next twenty years. We also determined that our aim was to do an identification and analysis of ethical issues associated with our subject, and we determined that we wanted to do a broad-scoped ethical analysis, not focusing on particular moral values or ethical issues, but on all major ethical issues associated with our subject of study.

In the second step, we engaged in creating thorough descriptions of our subject of study (i.e., present and future AI and robotics technologies, products and applications). These descriptions were based on consultation of AI and robotics experts and of literature in AI and robotics for the current state of the art, and foresight analyses for plausible future developments, for which we consulted AI and robotics experts and existing foresight studies. In the third step, we identified stakeholders and relevant (potential) uses and impacts associated with the technologies and applications of step 2, based on literature review, expert and stakeholder consultation, and additional foresight analysis.

In the fourth step, we identified present and potential future ethical issues with the technologies, products, applications and impacts that were identified in steps 2 and 3. These issues were identified based on a review of the ethics of AI and robotics literature, on expert and stakeholder consultation as well as on original ethical analysis that we performed ourselves. In step 5, finally, we analysed the ethical issues we identified in step 4, again basing ourselves on the existing ethics literature, stakeholder and expert consultation, and original ethical analysis. By analysis, we mean that we identified the moral values and principles that are at play in the moral issues that were identified, any potential conflicts between these values and principles, the roles, rights and interests of stakeholders, reasons or arguments for and against certain moral judgments, and the pros and cons of particular ways of addressing the value conflicts. In our study, we did not aim to make any final judgments about the rightness or wrongness of technologies, applications, uses or behaviours, or recommendations on how to proceed in the future.

For the sections on the overview and history of ethics of AI and robotics, and on the ethics of AI and robotics in different countries, we had special methodologies that are reviewed in our summaries of these sections.

Overview and history of ethics of AI and robotics

Section 3 of this contextualises the ethical analysis of sections 5 through 7 by providing brief histories of the ethics of AI and ethics robotics, covering both academic research and practical action, as well as by giving an overview of the present institutional support of these fields. The section is based primarily on literature analysis and an online search for important academic journals, academic conference series, and organisations and initiatives.

For the ethics of AI, we explained that the field has as its focus the ethical study of concepts, techniques and applications of AI, and that it has a degree of overlap with the ethics of robotics, to the extent that AI techniques are used in robots. We found that the field can be considered a constituent part of a broader *philosophy of AI*, which predates it, and that it has had only limited academic coverage before the 21st century. We detailed how at around 2005, the field received a big boost from early work in *machine ethics*, which theorises the implementation of moral decision-making faculties in computers and robots. Finally, we explained that since around 2015, there has been an explosion of publications in the ethics of AI discussing ethical issues ranging from concerns about algorithmic bias and human rights to concerns about transparency, explainability in AI and algorithmic accountability.

For the ethics of robotics, which is perhaps better known as *robot ethics* or *roboethics*, we explained that the field has focused on ethical aspects in the design, development, implementation, and treatment of robots. We listed some of the landmark events that propelled the field forward, including the *First International Symposium of Roboethics* in 2004, and noted the importance of *Roboethics Atelier Project*, which set out to design the first *Roboethics Roadmap*, giving further direction to the field. Finally, we explained that the academic debate has focused on a broad range of ethical issues, which include potential harms to autonomy, dignity, and privacy, and unemployment, moral responsibility, and overall wellbeing, and that there have been very critical appraisals within the roboethics community of the development and use of lethal robots for military and police purposes.

As for the present institutional support of fields of ethics of AI and robotics, we listed some of the most important academic journals, academic conference series, and organisations and initiatives that exist within these fields.

Ethics of AI and robotics in different countries

Section 4 of this report presents the results of a study that we conducted of how ethical issues in AI and robotics have been debated in different countries, both in the EU and globally, and to identify differences and similarities. The aim of this section is to provide context from national perspectives for the ethical analysis that follows, and also to provide building blocks for recommendations that we want to make later on in our project. Twelve countries were selected for our study, eight that are part of the EU (France, Germany, Poland, Sweden, The Netherlands, Greece, Spain, and the United Kingdom), and four other countries on different continents (United States, China, South Africa, and Brazil). We performed two related studies: (1) a study of national academic ethical discussions of AI and robotics, and (2) a study of national discussions of ethical, legal and social issues with AI and robotics in popular media. These studies were carried out by native experts: members of the SIENNA consortium with backgrounds in ethics or social science.

In our study of national academic ethical discussions of AI and robotics, we performed a search for, and analysis of the contents of, recent (2000–present) academic articles on the ethics of AI and robotics that had been authored by individuals from institutions within the country and were specifically addressing the situation within the country. We did so using relevant keywords in Google Scholar. In some countries, we observed broad coverage of ethical issues in AI and robotics (China, Germany, United States), whereas in others, it was more modest (France), and in still others, it was rather scant (Brazil, Greece, Poland, South Africa, Spain, Sweden, United Kingdom). The lack of country-specific ethics studies in the UK may be explained by the international academic orientation of UK institutions.

Across all twelve countries, the most widely discussed application areas of AI and robotics are defence, medicine, transportation, and the workplace, with the most-discussed products being autonomous weapon systems (especially “killer robots”), care robots, healthcare apps, surgical robots, sex robots, and autonomous vehicles. Especially notable was the significant amount of attention for the ethics of defence applications of AI and robotics in most countries. In most countries, a wide range of ethical issues were discussed, relating to justice, equality, autonomy, dignity, explainability, transparency, safety, accountability, liability, privacy, and data protection. This reflects the international academic debate. The most frequently mentioned issues were justice, privacy, and safety, which were often still addressed in countries were academic discussion was found to be scant. The national academic debates in the US, Germany and China stood out in also being focused on potential broad-scoped solutions to ethical issues, including through laws, standards, and regulation, as well as through ethics by design and implementation of moral reasoning systems in robots and AI systems.

In our study of national *popular media debates*, we performed a search, using relevant keywords in Google, for recent articles in national popular media on the ethical, legal and social issues in relation to AI and robotics and in relation to the country under study, and did an analysis of their contents. We observed that in all countries, with the possible exception of Poland, there has been substantial debate in the national popular media on ethical issues in relation to AI and robotics, although in some countries the debate has only recently gained pace. In most cases, the application areas, products, and ethical issues and principles addressed in the popular academic debate mirrored those in the academic debate. Issues related to the potential economic effects of AI and robotics technology, however, seemed to get slightly more attention.

General ethical issues in AI and robotics

In the first part of our ethical analysis of AI and robotics, we covered general ethical issues. These are ethical issues of three kinds: ethical issues associated with the general aims of AI and robotics, ethical issues with general techniques, methods and approaches in AI and robotics, and ethical issues with general implications and risks associated with the use of AI and robotics. In what follows, we summarize our results.

AI – general aims: We found that AI technology is being developed with the following aims in mind: efficiency and productivity improvement; effectiveness improvement; risk reduction; system autonomy; human-AI collaboration; mimicking human social behaviour; artificial general intelligence and superintelligence; and human cognitive enhancement. We then considered ethical critiques of each of these aims. We found, amongst others, that efficiency, productivity and effectiveness improvement are inherently tied to the replacement of human workers, which raises ethical issues. The mimicking of social behaviour is associated with risks of deception and of diminished human-to-human social interaction. The development of artificial general intelligence and superintelligence raises issues of human obsolescence and loss of control, and raises issues of AI and robot rights. Human cognitive enhancement, finally, comes with risks to equality, human psychology and identity, human dignity and privacy.

Robotics – general aims: For robot technology, we found the following general aims: efficiency and productivity improvement; effectiveness improvement; risk reduction; robot autonomy; social interaction; human-robot collaboration; novelty; and sustainability. Most of the ethical issues here mirror those with the aims of AI.

AI – techniques, methods and approaches: We identified the following general AI techniques and approaches and discussed associated ethical issues: algorithms; knowledge representation and reasoning techniques; automated planning and scheduling; machine learning; and machine ethics (i.e., the implementation of ethical decision-making capabilities in machines). For algorithms, we discussed how they can be value-laden and contain biases. In relation to knowledge representation, we discussed how inaccuracy, misrepresentation and bias can raise ethical issues. We discussed how automated scheduling and planning can raise issues of trustworthiness and responsibility, and could decrease human capabilities. In relation to machine learning, we discussed many ethical issues, including issues of transparency and explainability, fairness and discrimination, reliability, privacy and accountability. Machine ethics was analysed to have many pitfalls, including the difficulty of implementing human morality in AI systems, the potential for failure and corruptibility, equality of access to ethical AI, the undermining of human moral responsibility, and the possibility that we want to grant such systems moral status and rights.

Robotics – techniques, methods and approaches: We identified the following general AI techniques and approaches and discussed associated ethical issues: robot sensing, robot actuation, and robot control. For robot sensing, issues of reliability of error were discussed, as well as risks to privacy and safety associated with some sensor types. In relation to robot actuation, we discussed issues of safety, privacy, and psychological impacts. In relation to robot control systems, we discussed how robots can have different degrees of autonomy, and we discussed associated issues of safety, responsibility and accountability, transparency, and privacy.

AI – general implications and risks: We identified the following general implications and risks associated with the development and use of AI: potential negative implications for autonomy and liberty, privacy, justice and fairness, responsibility and accountability, safety and security, dual use and misuse, mass unemployment, transparency and explainability, meaningfulness, democracy and trust. (Sometimes, we also discussed potential positive implications.) For each value or issue, we aimed to come to a precise determination of it, we then discussed different general ways in which AI might impact it, and we analysed the moral considerations involved.

Robotics – general implications and risks: We identified the following general implications and risks associated with the development and use of robots: loss of control, autonomy, privacy, safety and security, dual use and misuse, mass unemployment, human obsolescence, human mistreatment, robot rights, and responsibility and accountability. We analysed these issues like we did in the corresponding section on AI.

Ethical issues concerning AI and robotics products

In this second part of our ethical analysis, we covered ethical issues with specific products, systems and processes in AI and robotics.

AI – products: For AI, we identified seven types of AI systems and subsystems that raise important or unique ethical concerns. They are intelligent agents, knowledge-based systems, computer vision systems, natural language processing systems, affective computing systems, (big) data analytics systems, and embedded AI & Internet of Things.

Intelligent agents are software programs that can autonomously enact goals in an environment. Ethical issues with them include privacy, user autonomy, trust, moral responsibility and liability, and questions about how ethical behaviour is best instilled in these constructs. Knowledge-based systems are computer programs that use a knowledge base to draw inferences and solve complex problems. Ethical issues include bias in knowledge representation and inferential patterns, self-modification of such systems that leads to unpredictable outcomes, accuracy, and security. Computer vision systems raise ethical concerns in relation to object detection, image classification, object recognition, and visual biometric applications (such as face, iris and fingerprint identification). They raise concerns about security, accuracy, privacy, and the expanded monitoring and surveillance capabilities that they offer.

Natural language processing systems raise issues of privacy (e.g., for speech processed by consumer systems like Siri, Amazon Echo, and Google Home, but also for online written text that can be analysed), and potential bias and discrimination in algorithms and use of data. Affective computing systems are systems capable of detecting, recognizing, interpreting, simulating and responding to emotions. They raise significant issues of privacy and trust, issues with using affective capabilities for deception, and unwanted social bonding and loss of autonomy. (Big) Data analytics systems, that are often used to process vast amounts of personal information, raise major issues of individual and group privacy, potential algorithmic bias and discrimination, and issues of transparency and accountability. Embedded AI & Internet-of-Things, finally, concerns AI embedded in electronic devices like vacuum cleaners and washing machines, and the networking of such devices in what has been called the Internet-of-Things. These devices raise serious issues of privacy, security and trust, since much personal information is sent between them, and it is often possible for them to get hacked. There are also concerns with devices actively limiting the autonomy and freedom of users and third parties, and the technology raises accountability issues.

Robotics – products: For robotics, we identified ten types of robotic systems that raise important or unique ethical concerns. They are humanoid robots, social robots, unmanned aerial vehicles, self-driving vehicles, telerobotic systems, robotic exoskeletons, biohybrid robots, swarm robots, microrobots, and collaborative robots.

Humanoid robots, robots that look and behave like humans, could easily become the subject of misplaced moral accountability, misplace trust, and misplaced empathy. They could be mistaken for real human beings by children and people with cognitive impairments, and could also reinforce stereotypes and be used to perpetuate socially undesirable behaviour. Social robots, robots designed to interact with humans through social behaviour, raise many of the same ethical issues as humanoid robots. They also raise the broader question of the context in which they should or should not be used, such as uses that substitute for human-human interactions in schools, healthcare, or home life, and uses by members of vulnerable groups. Unmanned aerial vehicles, or drones, raise issues of privacy, accountability, security, and transparency, and more generally the uses to which they should be put. Should we allow, for example, drones that are armed (for law enforcement use)? Where should drones be able to fly and make recordings? Self-driving or autonomous vehicles also raise issues of privacy, accountability, security and transparency, and raise ethical issues concerning the implemented crash algorithms and the way in which they make decisions in general.

Telerobotic systems, which are semi-autonomous robots operated from a distance, raise issues in terms of diminished social interaction between humans, negative effects on the psychological well-being of operators, and specific harms from increased technologisation, as well as issues of safety, security, equality, and responsibility. Robotic exoskeletons, which are wearable robots, raise issues of possible negative physical and psychological impacts on users, issues of access and equality, privacy, safety, and security, and the possibility of dehumanisation or overworking of industrial labourers. Biohybrid robots include both robots that include organically grown components and robots that imitate functions of organic lifeforms. The first type, especially, raises ethical issues concerning moral status and permissibility. Swarm robots, collections of often small and adaptive robots capable of collective decision-making, raise concerns because of their great potential for surveillance, and their potential unpredictability and uncontrollability. Safety and security are also a concern, as are their potential military applications. Microrobots, which are small and cheap robots that are used to access hard-to-reach areas, raise issues of surveillance and privacy, control and ownership, safety, and environmental degradation. Collaborative robots, finally, are robots designed to perform tasks in tandem with human labourers, for example in construction or medical intervention. They raise serious issues of trust and risks of psychological harm for human co-workers, and issues of privacy and security.

Ethical issues in different application domains

In this third part of our ethical analysis, we covered ethical issues with the application of AI and robotics in different application domains, such as healthcare, education, and defence, as well as ethical issues for different types of AI and robotics users and stakeholders.

AI – application domains: We identified thirteen major application domains for AI that raise important or unique ethical concerns. They are infrastructure and cities, healthcare, finance and insurance, defence, law enforcement, the legal sector, public services and governance, retail and marketing, media and entertainment, smart home and companionship, education and science, manufacturing, and agriculture.

Frequently recurring ethical issues in these different domains are privacy, transparency, responsibility, fairness, freedom, autonomy, security and trust. For domains in which they are an issue, we discuss their particular manifestations and peculiarities. Healthcare applications of AI raise special issues regarding potential risks to privacy and trust, threats to informed consent, discrimination, and risks of further increasing already existing health inequalities. Law enforcement applications raise issues of bias and discrimination, surveillance, and the risk of a lack of accountability and transparency for law enforcement decisions. Defence applications come with possible negative effects of AI on compliance with the principles of just war and the law of armed conflict, the possibility for uncontrolled or inexplicable escalation, and the potential for responsibility gaps.

In media and entertainment, we discussed ethical issues in news media, social media and audio and visual media. In news media, there is the risk of impoverished journalism, hyper-personalization that contributes to “filter bubbles”, and smart generation of fake news. In audio and visual media, like film and music, AI could undermine creativity if pushed too far, instituting formulaic processes that lack the creativity, spontaneity and humanity that human creators can bring. In social media, harvesting of personal information for advertising and political microtargeting could undermine privacy and democracy, AI could stimulate the formation of “echo chambers”, and there are controversies around automated social media censorship. AI in the agricultural sector could further increase the power imbalance between agribusinesses and farmers, and could reinforce big industrial monocultures. Other application domains also raise various unique issues.

Robotics – application domains: We identified ten major application domains for robotics that raise important or unique ethical concerns. They are transportation, law enforcement, defence, infrastructure, healthcare, companionship, manufacturing, exploration, service sector, and environment and agriculture.

Frequently recurring ethical issues in these domains are privacy, transparency, responsibility, fairness, autonomy, safety and trust. For domains in which they are an issue, we discuss their particular manifestations and peculiarities. Transportation applications, involving automated vehicles, raise significant issues, of trust, accountability, transparency, security and safety, which we explore. In healthcare, the application of care robots and surgical robots raises issues of accountability, patient privacy and confidentiality, maintenance of quality of care and patient integrity, and the risks of reduced humanity in patient care.

The topic of companionship covers applications of companion robots, such as robot pets, robot nannies, conversational robots and sex robots. Ethical issues include security, privacy and safety, possible negative implications for human-human interaction, and the appropriateness of certain applications of companion robots, for example for child care, elderly care, and sex and romantic relationships. In the service sector, including retail, recreation, restaurants, banking, and communications, amongst others, an issue is the extent to which robots should be able to make decisions without human approval or interference, and the value trade-offs this involves. Two other issues concern the replacement of human workers by service robots, and the risk of resemblances to slavery in certain service robot applications. The other mentioned application domains also raise various special ethical issues.

AI and robotics – issues for different types users and stakeholders: We identified and discussed ethical issues that concern different types of (vulnerable) end users and other stakeholders of AI and robotics technologies. We considered the following demographic categories: gender, race and ethnicity, age (with a focus on children and the elderly), ability (with a focus on people with mental and physical disabilities), educational level, and income level. With respect to gender, ethical issues include the possibility of women being disproportionally affected by AI-induced unemployment, algorithmic and functional gender bias and gender stereotyping in the design of AI and robotics products (to the detriment of women), and the lack of women in the AI and robotics technology sectors. With regard to race and ethnicity, ethical issues include algorithmic racial bias in the design of AI products, and humanoid robots contributing to the perception of particular racial groups in society as slaves. With respect to children, ethical issues include the shaping of children's views by biased AI systems and robots, a potential loss of social interaction with other children, stunted empathy development in children, and potential harms to privacy by intelligent Internet-connected toys.

With regard to the elderly, ethical issues include potential harms to privacy, the generation of false expectations about the (social) abilities of anthropomorphic robots, the potential for patronisation of elderly individuals by robots, and a potential loss of social interaction with other human beings. With regard to people with physical and mental disabilities, ethical issues include risks of dependency on AI systems and robots and increased social isolation, a diminished perception of social responsibility among human caregivers, and distributive justice concerns. With respect to educational and income level, ethical issues include unequal effects of AI and robotics on people depending on their level of education, and increased inequalities between the developed world and the developing world.

Conclusion

We have summarised the content of the SIENNA report on ethical analysis of artificial intelligence and robotics. We reviewed the objectives and structure of the report, reviewed its methodology, and summarized its major findings: those concerning past academic and practical activity in ethics of AI and robotics, those of a study of academic and popular discourses on ethical aspects of AI and robotics in various EU and non-EU countries, and those of current and potential future ethical issues with AI and robotics, including both general issues, issues relating to particular types of products, and issues relating to particular application domains.

This report can be read as a stand-alone report, but is part of a larger project on ethical and human rights aspects of emerging technologies. Other deliverables of the SIENNA project can be found on its website, at the following address: http://www.sienna-project.eu/publications/deliverable-reports/. Inquiries regarding this report can be directed at the two lead authors.

List of tables

List of figures

List of acronyms and abbreviations

Abbreviation	Explanation
AI	Artificial intelligence
ANN	Artificial neural networks
AUV	Autonomous underwater vehicle
BEAM	Biology, electronics, aesthetics and mechanics
CAD	Computer aided design
Cobot	Collaborative robot
EC	European Commission
GPS	Global Positioning System
IoT	Internet of Things
ITS	Intelligent tutoring system

MAV	Micro aerial vehicle
MEMS	Microelectromechanical system
NLP	Natural language processing
NPC	Non-player character
R&D	Research and development
SAR	Socially assistive robot
SEIA	Socio-economic impact assessment
UAV	Unmanned aerial vehicle

Table 1: List of acronyms/abbreviations.

Glossary of terms

Term	Explanation
Actuator	A device module or subsystem for performing actions in an environment.
Algorithm	"[A] precisely-defined sequence of rules telling how to produce specified output information from given input information in a finite number of steps."[1]
Artificial intelligence	The science and engineering of machines with capabilities that are considered intelligent (i.e., intelligent by the standard of *human* intelligence).
Artificial neural network	An interconnected network of simple and often uniform units similar to those that exist in the biological brain, which can be implemented in intelligent computing systems.
Autonomy	"[A] capacity to operate in a real-world environment without any form of external control, once the machine is activated and at least in some areas of operation, for extended periods of time."[2]
Big data	Extremely voluminous data sets that require specialist computational methods to uncover patterns, associations and trends in them.
Computer vision	An application of AI that gives a computer system the capacity to acquire, process and analyse (numerical or symbolic) information about the content presented in digital imagery.
Connectionist AI	A group of methods in AI research that utilise interconnected networks of simple and often uniform units similar to those that exist in the biological brain.
Data mining	The process of discovering patterns in large data sets involving database systems, statistical analysis, and intelligent methods such as machine learning.

[1] Knuth. Donald. "Computer Science and Its Relation to Mathematics," *American Mathematical Monthly*, Vol. 81, No. 4, 1974, pp. 323-343.

[2] Lin, Patrick, Keith Abney and George A. Bekey, "Current Trends in Robotics: Technology and Ethics," Robot Ethics: The Ethical and Social Implications of Robotics, MIT Press, 2012.

Deep learning	An approach to machine learning that applies artificial neural networks with hidden layers and the backpropagation method, in combination with powerful computer systems and voluminous training data.
Drone	Synonymous with "unmanned aerial vehicle"; an aircraft without a human pilot aboard.
Expert system	A computer system that can mimic a human expert's decision-making ability within a particular field by reasoning through a large amount of field-specific knowledge contained in a database.
Humanoid robot	A robot that resembles a human being in terms of appearance and/or behaviour.
Impact	A potential change – whether positive or negative, direct or indirect, in whole or in part – caused by or associated with the technological field under consideration.
Intelligence	A general cognitive ability encompassing several more specific abilities, including the abilities to reason, solve problems, plan, conceptualise, use language, and learn.
Intelligent agent	An artificially created, autonomous entity that can perceive its environment by means of sensors, act upon this environment through the use of actuators, and direct its activities towards reaching goals.
Internet of Things (IoT)	The interconnection via the Internet of objects in the physical world – devices, vehicles, persons, buildings and other items – allowing them to send and receive data.
Machine learning	A set of approaches within AI where statistical techniques and data are used to "teach" computer systems how to perform particular tasks, without these systems being explicitly programmed to do so.
Natural language processing	An application of AI that gives a computer system the capacity to understand human language in written or spoken form.
Robot control system	A system that uses a robot's sensor data to calculate and send appropriate signals to the robot's actuators.
Robotics	The field of science and engineering that deals with the design, construction, operation, and application of robots.
Robot	Electro-mechanical machines with sensors and actuators that can move, either entirely or a part of their construction, within their environment and perform intended tasks autonomously or semi-autonomously.
Socio-economic impact assessment	The analysis used to identify and assess the social, economic and environmental impacts of AI and robotics on society.
Sensor	A device, module or subsystem for detecting (and sending information about) events or changes in an environment.
Social robot	A robot that is capable of interacting with humans through social behaviour and adherence to rules attached to their social role.
Symbolic AI	A group of methods in AI research that are based on high-level, "symbolic" representations of problems, concepts, objects, events, etc.

Table 2: Glossary of terms.

1. Introduction

This SIENNA deliverable offers a broad ethical analysis of artificial intelligence (AI) and robotics technologies. More specifically, it identifies and analyses the present and potential future ethical issues in relation to: (1) the AI and robotics subfields, techniques, approaches and methods; (2) their physical technological products and procedures that are designed for practical applications; and (3) the particular uses and applications of these products and procedures. This deliverable also provides an overview of the history and state of the art of the academic debate on ethics of AI and robot ethics, as well as an overview of the current institutional support of these fields. Furthermore, the report presents a summary of our "country studies" analyses of the national academic and popular media debate on the ethical issues in AI and robotics in twelve different EU and non-EU countries.

Objectives

The primary aims of this report have been to comprehensively *identify* and further *analyse* the most important present and potential future ethical issues in relation to AI and robotics technology, their products, and their applications. In our ethical analysis, we strove to provide ample clarification, details about nuances, and contextualisation of the ethical issues that were identified, while avoiding the making of moral judgments and proposing of solutions to these issues.

A secondary aim of this report has been to convey the results of SIENNA's "country studies" of the national academic and popular media debate on the ethical issues in AI and robotics in twelve different EU and non-EU countries, highlighting the similarities and differences between these countries. In comparison to the aforementioned methods, our analysis of the country study results has contributed fairly little to the overall identification and analysis of the ethical issues in this report. However, the country study results are expected to contribute more significantly to future SIENNA deliverables.

Relation to other SIENNA work

This report follows previous SIENNA work on describing the state of the art of the fields of AI and robotics. *SIENNA D4.1 – State-of-the-art review of AI and robotics*, which was published in July of 2019, offers a detailed analysis of both fields in terms of their central concepts, their history, their present and anticipated technologies and applications, as well as a socio-economic impact assessment (of present and expected impacts) of their technologies.[3] Our analysis in this state-of-the-art review is based on a thorough literature review and commentary on our work by field experts.

Concurrent with writing the present report, the SIENNA consortium has planned, conducted and analysed citizen surveys in thirteen EU and non-EU countries, as well as citizen panels in five EU countries, both of which were aimed at obtaining insight into public awareness of and public opinions about present and future developments in AI and robotics. Two reports on this—one regarding the panels and one on the surveys—have been submitted to the European Commission alongside this report.

The present report lays the groundwork for a number of future SIENNA reports. Most importantly, analysis in terms the moral valence of the ethical issues that have been identified and described here

[3] Jansen, Philip, Stearns Broadhead, Rowena Rodrigues, David Wright, Philip Brey, Alice Fox, Ning Wang, *SIENNA D4.1 State-of-the-art Review*, WP4 - AI & Robotics, 2018, Public deliverable report from the SIENNA project. http://www.sienna-project.eu/publications/deliverable-reports/

will follow in the following future deliverables: *SIENNA D4.7 – Proposal for an ethical framework for AI and robotics*; and *SIENNA D5.4 – Central elements of a code of responsible conduct for researchers relating to AI and robotics*.

Definitions, scope and limitations

This report makes use of the same definitions of AI and robotics that have been advocated in the aforementioned SIENNA D4.1 report. AI can be defined as "the science and engineering of machines with capabilities that are considered intelligent, that is, intelligent by the standard of *human* intelligence." And robotics can be defined as "the science and engineering of programmable electro-mechanical machines that can perform human tasks autonomously or semi-autonomously." *(For more detailed definitions and descriptions of AI and robotics, please see our D4.1 report.)*

As is apparent from these definitions, there exists a degree of overlap between AI and robotics. Artificially intelligent machines may or may not be physically embodied and (semi-)autonomous (i.e., they are robots); and robots may or may not use AI techniques as a part of their control systems. The fields of AI and robotics come together in the science and engineering of *artificially intelligent robots*. We discuss the ethical issues in relation to such robots in the parts of this report that are focused on the ethical issues in robotics (subsections 5.2, 6.2 and 7.2).

Two important notes regarding the scope of our work need to be made. First, in order to provide the most useful input for the development of practical recommendations in later SIENNA reports, it has been deemed helpful to set a limit on the inclusion of potential developments in AI and robotics that may only occur over larger time scales. In the analysis of ethical issues relating to potential future developments in AI and robotics, we therefore have restricted ourselves to discussing developments that are reasonably possible within approximately twenty years from now, with most emphasis put on developments five to ten years from now. We consider a time horizon of twenty years to be neither a point in time too far into the future (making the analysis too speculative), nor one that is too close to the present (decreasing the anticipatory value of the analysis).

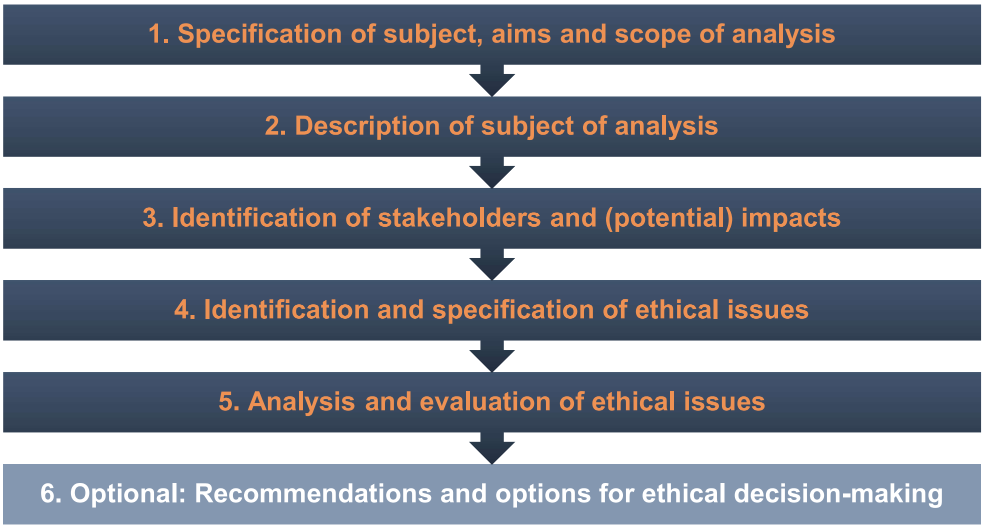
Second, as has been indicated earlier, this report is intended to form the groundwork for further SIENNA work on the moral valence of the issues that have been identified and analysed. As such, it provides no moral judgments regarding the goodness or rightness of particular actions, persons, things and events, and the rightness or wrongness of possible courses of action in relation to the ethical issues that have been identified. In the upcoming SIENNA report D4.7, considered moral judgments will be made for the ethical issues analysed here so as to arrive at an ethical framework for AI and robotics.

Structure of the report

The remainder of this deliverable is structured as follows. In section 2, we provide an overview of the history and state of the art of ethics of AI and robot ethics, as well as an overview of the current institutional support of these fields. In section 3, we give further details on the ethical analysis methodology that was used in this report. In section 4, we present a summary of our analyses of the national academic and popular media debate on the ethical issues in AI and robotics in twelve different EU and non-EU countries. In section 5, we identify and analyse the main ethical issues with regard to AI and robotics at the technology level of ethical analysis. In section 6, we identify and analyse the main ethical issues with regard to AI and robotics at the level of products and procedures. In section 7, we identify and analyse the main ethical issues with regard to AI and robotics applications. Finally, in section 8, we conclude with a summary and recommendations for further study.

2. Methodology for ethical analysis of AI and robotics

This section describes the methodology that has been used for the ethical analysis of AI and robotics technologies in this report (sections 5, 6 and 7). Previously, SIENNA researchers developed a methodological approach for ethical analysis in the project that can be found in *SIENNA D1.1 – The consortium's methodological handbook*.[4] The approach consists of a six-step process that is visualised in figure 1. For the current report, we carried out steps 1 through 4, and part of step 5. The sixth step, recommendations and options for ethical decision-making, will be carried out in a later report. The remainder of this section details our application of the first five steps.



1. Specification of subject, aims and scope of analysis
2. Description of subject of analysis
3. Identification of stakeholders and (potential) impacts
4. Identification and specification of ethical issues
5. Analysis and evaluation of ethical issues
6. Optional: Recommendations and options for ethical decision-making

Figure 2: Overview of the SIENNA approach to ethical analysis.

Step 1: Specification of subject, aims and scope of ethical analysis

In the first step of writing this report, we specified the subject, aim and scope of the ethical analysis to be performed. We began by identifying and defining the technologies, technological products, and application domains that we wanted to study: AI and robotics technology and their various manifestations and applications, both at present, and as they may evolve in the future. We then determined that our aim for the ethical analysis was to do an identification and analysis of ethical issues associated with our subject, and we determined that we wanted to do an ethical analysis with broad scope, not focusing on particular moral values or ethical issues, but on all major ethical issues associated with our subject of study. We also determined that we would not perform ethical evaluations of the ethical issues we analysed, meaning that in this report we would not arrive at considered moral judgments about these issues.

With regard to the analysis of potential future ethical issues associated with our subject, we decided to limit our scope to those issues that can potentially occur between now and twenty years into the future, with special emphasis put on issues that have a reasonable likelihood of occurring within five

[4] Rodrigues, Rowena, et. al., *D1.1: The consortium's methodological handbook*, WP1, 2018, Public deliverable report from the SIENNA project.

to ten years from now. A time horizon of 20 years was chosen since it was considered neither a point in time too far into the future (making the analysis too speculative), nor one too close to the present (decreasing its anticipatory value).

Step 2: Description of subject of analysis

In the second step, we engaged in thoroughly describing of our subject of study. To perform a broad-scoped ethical analysis, we needed broad descriptions of our subject that included different AI and robotics subfields, techniques, produced artefacts and uses, both present ones and ones that may take place in the future. Following the *Anticipatory Technology Ethics* approach developed by Brey (2012),[5] we structured these descriptions along three "levels of description": (1) the *technology level*, the most general level of description, which specifies the technology in general, its subfields, and its fundamental techniques, methods and approaches; (2) the *artefact level* or *product level*, which provides a systematic description of the technological artefacts (physical entities) and procedures (for achieving practical aims) that are being developed on the basis of the technology; and (3) the *application level*, which defines particular uses of these artefacts and procedures in particular contexts by particular users.

Methods for making descriptions of our subject of analysis at the three levels of description have included: (1) literature review and expert consultation (the latter through workshops and interviews[6]) to obtain insights into current state of the art in the fields of AI and robotics; and (2) foresight analysis through expert consultation and analysis of existing foresight studies to obtain insights into plausible future developments in these fields.

It should be noted that, prior to writing this report, much (though not all) of the work in this step had already been conducted for an earlier SIENNA report: *SIENNA D4.1 – State of the art review of AI and robotics*.[7]

Step 3: Identification of stakeholders and (potential) impacts

In the third step, we specified current and potential future impacts associated with our subject of ethical analysis, focussing on social, economic, environmental, and other kinds of impacts at the micro, meso and macro levels. We identified these impacts in relation to the three levels of description outlined in step 2: (1) broad impacts correlated with the technology in general and its core fields and techniques; (2) impacts correlated with specific artefacts; and (3) impacts correlated with specific uses.

Methods used to specify the impacts have included literature review (of the socio-economic impact assessment literature), expert and stakeholder consultation, and additional foresight analysis. Much of the work on specifying the impacts had already been conducted as a part of the SIENNA D4.1 report.[8]

[5] Brey, Philip, "Anticipatory Ethics for Emerging Technologies," *Nanoethics*, Vol. 6, 2012, pp. 1–13.

[6] At the end of 2018, we have conducted a small workshop and nine one-on-one interviews with AI and robotics technology experts that have helped us to obtain insights into the current state of the art of AI and robotics and plausible future developments in these fields. We are grateful to Professor Joanna Bryson for her contribution.

[7] Jansen, Philip, Stearns Broadhead, Rowena Rodrigues, David Wright, Philip Brey, Alice Fox, Ning Wang, *SIENNA D4.1 State-of-the-art Review*, WP4 - AI & Robotics, 2018, Public deliverable report from the SIENNA project. http://www.sienna-project.eu/publications/deliverable-reports/

[8] Ibid.

Furthermore, in this step, we identified and specified relevant stakeholders (e.g., decision makers, those involved in benefitting or being harmed by the subject or its impacts) and made plans to engage them. For the current report, stakeholders have mostly been engaged through our SIENNA workshops.

Step 4: Identification and specification of potential ethical issues

In the fourth step, we identified and described all present and potential future ethical issues regarding (and all principles and values that may be affected or challenged by) the AI and robotics technologies, products, applications and impacts that were described during steps 2 and 3.

In identifying and describing the ethical issues, we again followed Brey's *Anticipatory Technology Ethics* approach by using three "levels of ethical analysis": the technology level, the artefact level (or product level), and the application level. At the technology level, we identified (1) ethical issues regarding the aims of AI and robotics research and development, (2) ethical issues with respect to the central concepts, subfields, techniques, methods, and approaches used in AI and robotics, and (3) general ethical issues that apply to most or all AI and robotics products and applications and their impacts on society. At the artefact level, we identified ethical issues that typically occur for certain types of AI and robotics products or procedures across a wide range of applications of them. And at the application level, we identified ethical issues with respect to the technology and its specific products (1) in specific application domains (e.g., healthcare, defence, domestic use), (2) in non-western countries, and (3) in use by specific types of users (e.g., children, the elderly, women, people with disabilities). Table 3 below provides an overview of the central questions of ethical analysis for each of the three levels.

Level of analysis	Objects of analysis	Questions for ethical analysis
Technology level	- Aims of the technological field - Broad features of the technological field (central concepts, methods, approaches) - General features and impacts that apply to artefacts and applications emerging from the field	- What are ethical issues, if any, regarding the aims of the field, or of particular subfields, methods and approaches? - What are ethical issues, if any, regarding central concepts, methods, subfields, and approaches in the field? - What are general ethical issues that apply to most or all artefacts and applications coming out of the field and their impacts on society?
Artefact level	- Technological artefacts (products) - Technological procedures (functional procedures developed within the field) (Both developed for use outside the field)	- What ethical issues (typically) occur for certain types of products or procedures (across a wide range of applications of them)?
Application level	- Uses of technological artefacts/procedures in particular domains or contexts, for particular purposes or by particular user groups	- What ethical issues occur with respect to the technology and its specific products in healthcare, defence, domestic use, etc., in non-western countries, in use by children, the elderly, men, people with disabilities, etc.?

Table 3: Overview of objects and central questions for ethical analysis of SIENNA's approach to ethical analysis.

In this report, the three levels of ethical analysis are each covered in a separate section: section 5 for the technology level, section 6 for the product level, and section 7 for the application level.

Methods for the identification and specification present and potential future ethical issues at the three levels of analysis have included: (1) literature review of prior ethics studies in the fields of AI and robotics, (2) stakeholder and expert consultation through workshops and interviews,[9] and (3) the use list of questions about the technologies that could help identify ethical issues (which are sometimes presented as "checklists"[10]), e.g., by cross-referencing them with the results of our SIENNA D4.1 report on the state of the art of AI and robotics technology.

Step 5: Analysis of ethical issues

Having had identified the ethical issues in relation to AI and robotics technologies, the final step in writing this report was to try to better understand and further analyse these issues. This involved steps to further clarify, provide details about nuances, and contextualise the ethical issues that were identified. These steps included identifying different moral values and principles that are at play in the issues and potential conflicts between these values and principles, as well as identifying roles, rights and interests of stakeholders.

Note that in this report, we have only partially executed step 5 of the SIENNA handbook's approach to ethical analysis: Our analysis has not focussed on providing ethical *evaluations* of the issues that have been identified or on suggesting ways to solve them. This means that we have not made moral judgments regarding the goodness or rightness of particular actions, persons, things and events, and the rightness or wrongness of possible courses of action in relation to the ethical issues that have been identified. In the upcoming SIENNA D4.7 report, considered moral judgments will be made for the ethical issues analysed here so as to arrive at an ethical framework for AI and robotics.

As with the previous step, the results of this analysis step have also been structured along the three levels of ethical analysis provided by Brey's *Anticipatory Technology Ethics* approach. Accordingly, the analysis of identified ethical issues at the technology level is covered in section 5; the analysis of issues at the product level is provided in section 6; and the analysis of issues at the application level is given in section 7.

Methods for the analysis of identified present and potential future ethical issues have included: (1) literature review of studies that conduct in-depth analysis of ethical issues in AI and robotics, (2) expert consultation through workshops,[11] and (1) original ethical analysis through application of instruments from the field of ethics (i.e., ethical concepts, theories, frameworks and/or arguments).

[9] In January of 2019, we organised a two-day workshop in London on the identification of present and future ethical issues in AI and robotics that was attended by around 20 stakeholders, ethicists and technology experts. The results of this workshop are reflected in the report.

[10] Several ethical checklists are available. Brey, op. cit., 2012 contains a comprehensive checklist for ethical issues in technology, and the SATORI CEN "pre-standard" for ethics assessment also specifies a large number of ethical issues in relation to the medicine, information technology and engineering fields. See: SATORI, "CEN Workshop Agreement: Ethics assessment for research and innovation - Part 2: Ethical impact assessment framework, CWA 17145-2, June 2017. http://satoriproject.eu/media/CWA17145-23d2017.pdf

[11] In June of 2019, we organised a two-day workshop in Uppsala, Sweden, on the analysis of present and future ethical issues in AI and robotics, which was attended by around 20 experts in the ethics of AI and robotics. During this workshop, feedback was given on important parts of the ethical analysis sections of an earlier draft of this report.

3. Overview and history of the ethics of AI and robotics

This subsection offers a brief overview of the history of the ethics of AI and robotics, as well as an overview of the present institutional support of these fields. In doing so, it provides some context to the main ethical analysis parts of this report. Subsection 3.1 and 3.2, cover the history of the ethics of AI and history of the ethics of robotics, respectively. Subsection 3.3 covers the institutional support of these fields by listing some of the most important academic journals, academic conference series, and organisations and initiatives that exist for them.

3.1. History of ethics of AI

The ethics of AI has as its focus the ethical study concepts, techniques and applications of artificial intelligence. It has a degree of overlap with the ethics of robotics, to the extent that AI techniques are used in robots and give rise to ethical issues.

The field can be considered a constituent part of the broader *philosophy of AI*, which predates it. What is known today as the philosophy of AI emerged in the 1960s and became an established field in the 1980s.[12] The focus in this philosophical discipline has mainly been on assumptions and approaches within the scientific approach to AI, and its relation to cognitive science; notably less attention has been given to the engineering approach to AI.[13] The philosophy of AI considers questions such as whether machines (or more specifically computer systems) are capable of general intelligence, or whether they are capable of having mental states and consciousness. Questions are asked too about whether human intelligence and machine intelligence are essentially the same and if the mind therefore can be seen as a computational system. Philosophers have also explored the relation between philosophical logic and AI and ethical issues in AI (Section 4.6).[14]

The ethics of AI has had limited academic coverage before the 21st century. An important precursor to field, however, is Joseph Weizenbaum's monograph *Computer Power and Human Reason: From Judgment to Calculation*, which dates from 1976.[15] In this work, Weizenbaum conveys his ambivalence towards computer technology. His general message was that while AI may be possible, computers should never be allowed to make important decisions as they will always lack human qualities.

The relative lack of further scholarly attention before the turn of the century can be explained by the limitations in computing power and AI theory that existed at the time. As advances in these areas resulted in a renewed focus on the field of AI since the mid-2000s, however, the ethics of AI became a bona fide field of research.

The ethics of AI received a big boost from the emergence of work in *machine ethics*, a small field of research that gained traction with the AAAI Fall 2005 Symposium on Machine Ethics. Machine ethics

[12] Brey, Philip, and Johnny Søraker, "Philosophy of Computing and Information Technology," In *Philosophy of Technology and Engineering Sciences. Vol. 14 of the Handbook for Philosophy of Science*, (ed. A. Meijers) (gen. ed. D. Gabbay, P. Thagard and J. Woods), Elsevier, 2009.
[13] Ibid.
[14] Ibid.
[15] Weizenbaum, Joseph, *Computer Power and Human Reason: From Judgment to Calculation,* W. H. Freeman and Company.

(which is also known as *machine morality*, *artificial morality*, and *computational ethics*) sits at the intersection of ethics and computer science, and theorises the implementation of moral decision-making faculties in computers and robots. In other words, machine ethics aims to investigate ways to create machines that are guided by acceptable ethical principles in their decision making about the possible courses of action.[16] Two main reasons were identified for pursuing this area of inquiry. First, computational modelling of human morality was expected to help achieve a better understanding of human morality. And second, equipping machines with the capability to make decisions based on acceptable ethical principles was increasingly seen as indispensable requirement, given the increasing autonomy of machines and the fact that machines had been taking over more and more human tasks and operating in closer proximity to humans.

At present, machine ethics is clearly a subfield of the broader field of ethics of AI, as the latter has gained significant traction (although it is arguably still establishing itself). Since around 2015, there has been an explosion of publications in the ethics of AI discussing ethical issues ranging from concerns about algorithmic bias and human rights to concerns about transparency, explainability in AI and algorithmic accountability. An important development in recent years has been that computer science associations, IT companies and policymakers have acquired a strong interest in ethics of AI as part of their interest in AI in general as a key enabling technology.

One of the landmarks in the development of Ethics of AI has been the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, which has the goal of setting ethical standards for AI in the computer science, IT and electrical engineering fields. Many other initiatives, publications, organisations and conferences have emerged in recent years, the most important of which are listed in subsection 3.3.

3.2. History of ethics of robotics

The ethics of robotics, which is perhaps better known by the term *robot ethics* or *roboethics*, concerns ethical problems that occur with robots. More specifically, it focuses on the ethical aspects in the design, development, implementation, and treatment of robots.

An important event that propelled the field forward was the *First International Symposium of Roboethics* in San Remo in 2004. At this event, philosophers, ethicists, legal scholars, sociologists, anthropologists, together with robotic scientists, laid the foundations of ethics in the design, development and implementation of robots.[17] Apart from the symposium in San Remo, the *IEEE Robotics and Automation Society Roboethics Workshop: ICRA 2005* in Barcelona and the *Roboethics Mini-symposium: IEEE BioRob 2006 Biomedical Robotics and Biomechatronics Conference* in Pisa are seen as influential moments in the development of the field.[18]

In 2005, the European Robotics Research Network (EURON) funded the Roboethics Atelier Project, coordinated by the Scuola di Robotica.[19] The Atelier's first purpose was to produce a *Roboethics*

[16] Anderson, Anderson, ''IEEE Intelligent systems'', published by the IEEE Computer Society, 2006

[17] Veruggio, ''The Birth of Roboethics'', ICRA, IEEE International Conference on Robotics and Automation Workshop on Robo-Ethics, 2005

[18] Tzafestas, ''Roboethics: Fundamental Concepts and Future Prospects'', 2018

[19] Ibid.

Roadmap, a common tool for the interested community to (1) develop a common language on roboethics among scholars and stakeholders, and (2) learn about other fields, make connections and create new ideas.[20] The roadmap has provided a comprehensive review of the state of the art in the field of robotics and identified the major challenges for progress. The goal has been to identify the current driving forces, objectives, bottlenecks and key challenges for robotics (and robotics research), so as to develop a focus and guidance for the development of robotics in the next 20 years.

In the field of roboethics, like in the ethics of AI, the debates have only recently gained significant traction, but have focused on a broad range of ethical issues. They have included discussions on potential harms to autonomy, dignity, and privacy, and technological unemployment and the possible erosion of moral responsibility, which emerge through the design and application of service, social, industrial and other kinds of robots that interact with or affect humans in a variety of settings, such as healthcare, assisted living, and education. Finally, there have been generally very critical appraisals within the roboethics community of the development and use of lethal robots for military and police purposes.

3.3. Present institutional support for ethics of AI and robotics

In this subsection, we list some of the most important academic journals, academic conference series, and organisations and initiatives that currently exist within the fields of ethics of AI and ethics of robotics.

Academic journals

- *Ethics and Information Technology*[21]
- *Minds and Machines*[22]
- *AI & Society*[23]
- *Philosophy and Technology*[24]
- *Science and Engineering Ethics*[25]
- *International Journal of Social Robotics*[26]

Academic conference series

- *The Association for the Advancement of Artificial Intelligence (AAAI) and Association for Computing Machinery's (ACM) conference series on Artificial Intelligence, Ethics and Society (AIES)*[27]
- *The Robophilosophy conference series*[28]
- *The International Society for Ethics and Information Technology's (INSEIT) conference series on Computer Ethics Philosophical Enquiry (CEPE)*[29]

[20] Veruggio, ''The Birth of Roboethics'', ICRA, IEEE International Conference on Robotics and Automation Workshop on Robo-Ethics, 2005
[21] https://link.springer.com/journal/10676
[22] https://link.springer.com/journal/11023
[23] https://link.springer.com/journal/146
[24] https://link.springer.com/journal/13347
[25] https://link.springer.com/journal/11948
[26] https://link.springer.com/journal/12369
[27] http://www.aies-conference.com
[28] http://conferences.au.dk/robo-philosophy
[29] https://inseit.net/conferences/4

- *The International Association for Computing and Philosophy's (IACAP) conference series*[30]
- *The Society for Philosophy and Technology's (SPT) Biennual Meeting*[31]
- *ETHICOMP*[32]

Organisations and initiatives

- *International Society for Ethics and Information Technology*

  The International Society for Ethics and Information Technology (INSEIT) is a nonprofit (unincorporated) association that was created in 2000 to promote and facilitate scholarships, education, discussion, debate and other activities, on the ethical issues in and surrounded by information technology (IT).
  Link: https://inseit.net
- *The International Association for Computing and Philosophy*

  The International Association for Computing and Philosophy (IACAP) exists to promote scholarly dialogue and research on all aspects of the computational and informational turn, and on the use of information and communication technologies in the service of philosophy.
  Link: http://www.iacap.org
- *The Society for Philosophy and Technology*

  The Society for Philosophy and Technology (SPT) is an independent international organization that encourages, supports and facilitates philosophically significant considerations of technology.
  Link: http://www.spt.org
- *Institute of Electrical and Electronics Engineers' Global Initiative on Ethics of Autonomous and Intelligent Systems*

  The IEEE Global Initiative's mission is "to ensure every stakeholder involved in the design and development of autonomous and intelligent systems is educated, trained, and empowered to prioritize ethical considerations so that these technologies are advanced for the benefit of humanity."
  Link: https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html
- *High-Level Expert Group on Artificial Intelligence*

  The High-Level Expert Group on Artificial Intelligence (AI HLEG) of the European Commission has as its general objective to support the implementation of the European Strategy on Artificial Intelligence.
  Link: https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence
- *The Partnership on AI*

  Amazon, Google, Facebook, IBM, and Microsoft have established a non-profit partnership to formulate best practices on artificial intelligence technologies, advance the public's understanding, and to serve as a platform about artificial intelligence.
  Link: https://www.partnershiponai.org
- *Foundation for Responsible Robotics*

  The mission of the Foundation for Responsible Robotics (FRR) is to shape a future of responsible robotics and artificial intelligence (AI) design, development, use, regulation, and implementation.
  Link: https://responsiblerobotics.org
- *4TU.Centre for Ethics and Technology*

  4TU.Centre for Ethics and Technology is a community of researchers that aims to stimulate and perform research in the field of ethics and technology, both fundamental and applied.

[30] http://www.iacap.org
[31] http://www.spt.org
[32] https://www.unirioja.es/ethicomp

Link: https://ethicsandtechnology.eu

- *Algorithmic Justice League*

  The Algorithmic Justice League (AJL) is a collective started that aims to remove human bias from AI algorithms that can result in exclusionary experiences and discriminatory practices.
  Link: https://www.ajlunited.org

- *AI Now Institute*

  The AI Now Institute at New York University is an interdisciplinary research centre dedicated to understanding the social implications of artificial intelligence.
  Link: https://ainowinstitute.org

- *AI Ethics Lab*

  AI Ethics Lab brings together researchers and practitioners from various disciplines to detect and solve issues related to ethical design in AI.
  Link: http://aiethicslab.com

- *AI4ALL*

  AI4ALL is a non-profit working to increase diversity and inclusion in artificial intelligence.
  Link: http://ai-4-all.org

- *Open Roboethics Institute*

  Open Roboethics Institute (ORI) spun out of the Open Roboethics initiative, an international Roboethics think tank hosted at University of British Columbia.
  Link: https://www.openroboethics.org

- *Open AI*

  OpenAI is focused on discovering and enacting the path to safe artificial general intelligence. OpenAI conducts fundamental, long-term research toward the creation of safe AGI. The organization builds free software for training, benchmarking, and experimenting with AI.
  Link: https://openai.com

- *Future of Life Institute*

  The Future of Life Institute (FLI) is a charity and outreach organization working to ensure that tomorrow’s most powerful technologies are beneficial for humanity.
  Link: https://futureoflife.org

4. The ethical debate on AI and robotics in different countries

As explained in the introduction, a secondary aim of this report has been to convey the results of SIENNA's "country studies" of the national academic and popular media debate on the ethical issues of AI and robotics in twelve different EU and non-EU countries, highlighting the similarities and differences between these countries. These country study results have only been a minor contribution to the overall identification and analysis of the ethical issues in this report (sections 5, 6 and 7), but they are expected to contribute more significantly to future SIENNA deliverables that build on our research here.

In twelve different EU and non-EU countries, SIENNA partners have conducted limited studies in their institution's country of both academic literature and articles in the media on the topic of ethics of AI and robot ethics. We received completed studies from our partners in twelve countries: Brazil, China, France, Germany, Greece, the Netherlands, Poland, South Africa, Spain, Sweden, the United Kingdom, and the United States. The full reports of these studies are provided on the SIENNA website.[33]

In the remainder of this section, we describe the methodology of the SIENNA country studies (subsection 4.1), we present the summarised findings for each of the country studies (subsection 4.2), and we present a preliminary analysis of the findings, highlighting similarities and differences between the countries (subsection 4.3).

4.1. Methodology

For the "country study" task, SIENNA partners followed used a methodological approach outlined in this subsection. The task consisted of two parts: (1) a search for, and analysis of the contents of, recent (2000–present) *academic articles* on the ethics of AI and robotics that are specific to the country under study; and (2) a search for, and analysis of the contents of, recent (2000–present) *popular media articles* on the *ethical, legal and social issues* in relation to AI and robotics that are specific to the country under study. By "specific to the country under study" we mean that only those articles were included that had been authored by individuals at institutions within the country *and* were specifically addressing ethical issues with reference to the local context of the country (e.g., population, geography, economy, or other fundamental characteristics of the country).[34]

By academic articles we mean anything that can be found using Google Scholar, including academic journal articles, reports from government agencies/institutes, think-tanks and advisory organizations that are academically rigorous (i.e., contribute to the academic debate via interaction through standard citation formats). By popular media articles we mean newspaper articles, online news articles, popular science articles, weekly magazines about current affairs, books, et cetera, aimed at a broad and non-professional readership.

[33] Please see: http://www.sienna-project.eu/publications/deliverable-reports/

[34] A strong indicator for this criterion in non-English-speaking countries was the article being written in the national language of the country. The article could then at least identify the issues in the article as interesting for natives of the country.

The searches for academic articles were conducted through Google Scholar, and the searches for popular media articles were conducted through the regular Google search engine. Partners included a limited number of the most relevant articles in a detailed analysis of their ethics content (i.e., at most 20 articles each for the academic and popular media analyses, depending on the number of relevant articles they found).

The following search terms were suggested to the partners, which they could adapt to their country's situation (e.g., translate into the language of the country):

> For robotics: *("robots" OR "robotics" OR "automation" OR "automated" OR "machine" OR "machines" OR "unmanned" OR "driverless" OR "pilotless" OR "drones") AND ("ethics" OR "ethical" OR "social" OR "legal") AND <COUNTRY>*
>
> For AI: *("AI" OR "artificial intelligence" OR "intelligent agents" OR "automation" OR "smart systems" OR "big data") AND ("ethics" OR "ethical" OR "legal") AND <COUNTRY>*

Partners' analyses were standardised by means of a "reporting document" of around 10 pages, in which they were asked to list all the articles that they found and answer the following questions for each article:

- *What kinds of AI and/or robotics products, systems, or processes are discussed?*
- *What application areas are discussed?*
- *What ethical concepts, issues and values are discussed (state briefly)? And what is the expected timeline for these issues?*

In the same reporting document, partners were also asked to write summaries of their findings for both the academic analysis and the popular media analysis, in which they addressed the following questions:

- *Were the ethical, legal and social issues specific to your country?*
- *Can you contextualise these issues in the larger cultural, financial, religious, political or societal context of your country?*
- *Can you glimpse a trend based on years (2018–2013; 2012–2008, etc.)?*
- *Are there themes that are surprising to find, or surprising not to find?*
- *Did you find a preponderance on one issue and nothing on many others? Can you explain why this is?*

4.2. Summarized findings per country

This subsection offers brief summaries of the main findings for each that was country studied in terms of the ethical issues covered in the national academic and popular media debates. Please note that full summaries of the findings per country are available on the SIENNA website.[35]

4.2.1. Academic debate on ethical issues in AI

Country	Summarised findings

[35] Please see: http://www.sienna-project.eu/publications/deliverable-reports/

Brazil	About half of all articles found address issues related to the education of children. Most of the articles found were master's or PhD dissertations. It was noted that the articles were generally of a fairly low quality, with many drawing trivial conclusions.
China	In China, there has been significant attention on the ethics of the development and use of AI. Chinese scholars focus on specific ethical issues, safety and privacy in particular, but also responsibility, equality, justice, dignity, control. They also focus on specific concepts and theories in relation to AI, such as agency, subjectivity, responsible innovation, moral philosophy, moral algorithms, design ethics, and social ethics. Policy implications on the basis on these issues are also often discussed.
France	The issues that appear to be the most common in the French debate are those relating to the way in which the algorithms that lie at the basis of AI systems are formed. There are concerns that not AI may not only reproduce existing forms of injustice, while disguising itself with an aura of neutrality, but also that it could generating new and much stronger ones.
Germany	In Germany, there is substantial country-specific academic debate on ethical issues with AI. The ethical issues in relation to AI and robotics are mostly not discussed separately. The following AI and robotics applications and products are prominently discussed: applications in healthcare (care robots, healthcare apps, surgical robots), applications in transportation (especially driverless cars), applications in the workplace, and applications in defence. The most important issues relating to these are: privacy and data protection issues, responsibility and liability issues, changes in the workplace and unemployment, issues of safety, bias and discrimination (e.g., through facial recognition, algorithmic decision-making), issues of transparency, issues of control (automated systems dominating humans), and trust. Notably, on AI, there is discussion of how to train systems to act ethically and whether and how we can implement moral reasoning systems.
Greece	The academic discussion on ethical issues is scant and very recent (mostly after 2016). The articles address issues in the context of data protection, intellectual property, contract formation, and automated decision-making (algorithmic discrimination), and reference ethical principles such as privacy, autonomy, justice, safety, and control.
Netherlands	The academic debate about the ethical and social implications of AI in the Netherlands is not (yet) focussed on specific Dutch considerations of AI applications. Experts address universal issues for AI, which are discussed in the context of Dutch legal situation and policy making. Education and healthcare are fields for which current issues are addressed.
Poland	There has been little academic discussion of ethical issues specific to Poland. A significant part of the literature focuses on reviewing foreign literature and applying contexts to the Polish context. The articles focus mostly on issues of legal liability for AI, algorithmic transparency, bias and discrimination, and mass unemployment and the quality of work.
South Africa	There has been little academic discussion of ethical issues specific to South Africa. At most, and at a stretch, the articles reveal a concern that the introduction AI could further aggravate existing political and socio-economic inequalities (by promoting the health of the most well-off and relocation of jobs to high-income countries).
Spain	The academic discussion on ethical issues is rather scant. The main areas of focus in the Spanish academic debate are the military (AI in autonomous weapon systems), work and medicine. With regard to AI's effects on work, bias in algorithmic decision-making has been highlighted. With regard to medicine, there has been some debate about "the creation of life" and "playing God".
Sweden	The academic discussion on ethical issues with AI and robotics technologies is scant and very recent. All academic articles found were either students' master's or bachelor's theses, mostly in the areas of law and computer science. Ethical issues discussed in the articles relate to unemployment, worker safety, responsibility and liability, loss of control, privacy, intellectual property rights (for AI generated art), and "electronic personhood" for robots.

United Kingdom	The international and UK academic debate are so closely connected that they are hard to distinguish or detangle. It was hardly possible to identify a specifically "British" academic debate on ethics of AI and robotics. That being said, the few articles that were analysed mostly focused on issues relating to fairness, autonomy, transparency, accountability, privacy, data protection, consent, legitimate interest, governance, and compliance.
United States	The academic debate on ethical issues in AI and robotics in the US has focused, first and foremost, on drones and autonomous robots, especially drone warfare, autonomous weapons and the ethics of their use and design. The second most common topic (though a distant second) was applications of AI and robotics in the medical industry. The third most common topic was the study of human acceptance of AI & robotics broadly (whether attitudes toward robots in the home or of AI in the doctor's office, as just a few examples). Ethical issues discussed included or related to: justice, equity, explainability, transparency, acceptance, autonomy, safety, accountability, liability, privacy, data protection, consumer confidence, regulation, certification, laws of war, and rules of engagement, amongst others.

Table 4: Summarised findings per country on the country-specific academic debate on ethical issues in AI.

4.2.2. Academic debate on ethical issues in robotics

Country	Summarised findings
Brazil	About half of all articles found address issues related to the education of children. Most of the articles found were master's or PhD dissertations. It was noted that the articles were generally of a fairly low quality, with many drawing trivial conclusions.
China	In China, there has been significant attention on the ethics of the development and use of robotics. Some scholars focus on larger, more abstract and theoretical, themes related to robotics in machine ethics and robot ethics. Others focus more on issues in specific robot application areas, such as sex robots, medical robots, assistance robots, household robots, and autonomous vehicles. There is debate on ethical issues in relation to dignity, justice, safety, privacy, responsibility (especially the last three), and "harmonious relationships between humans and machines". As with research on the ethics of AI, policy implications are often also focused on.
France	The academic debate in France has four strands: One is about the impact that robots will have on work. A second one is about the nature and purpose of robots within our societies. A third one is about the risks of using robots and the necessity to foresee a regulatory framework, either internal (ethics) or external (laws). This last aspect raises the general question about dignity and the relation to our normative background that robots are already modifying. This discussion is often implicit in the analyses, but it is the main puzzling aspect that feeds the whole debate.
Germany	In Germany, there is substantial country-specific academic debate on ethical issues with AI. The ethical issues in relation to AI and robotics are mostly not discussed separately. The following AI and robotics applications and products are prominently discussed: applications in healthcare (care robots, healthcare apps, surgical robots), applications in transportation (especially driverless cars), applications in the workplace, and applications in defence. The most important issues relating to these are: privacy and data protection issues, responsibility and liability issues, changes in the workplace and unemployment, issues of safety, bias and discrimination (e.g., through facial recognition, algorithmic decision-making), issues of transparency, issues of control (automated systems dominating humans), and trust.
Greece	The academic discussion on ethical issues is scant and very recent (mostly after 2016). The articles address issues in the context of data protection (including drones), intellectual property, and contract formation, and reference ethical principles such as privacy, autonomy, justice, safety, and control.

Netherlands	Striking is the number of articles about the ELSI of robotics in healthcare. Besides considerations in healthcare, opportunities and concerns about robotics in education and the labour market are discussed. Most often these articles reflect on 'good care' and the trade-off between autonomy and improvement of well-being by use of a care robot. Questions with respect to labour are mainly focussed on employment and responsibility.
Poland	There has been little academic discussion of ethical issues specific to Poland. A significant part of the literature focuses on reviewing foreign literature and applying contexts to the Polish context. The primary focus of the articles is on issues of mass unemployment, quality of work, and safety. In addition, there is limited discussion of application-specific issues relating to the use of robots in the military and sex robots.
South Africa	There has been little academic discussion of issues specific to South Africa. At most, and at a stretch, the articles reveal a concern that the introduction robotics could further aggravate existing political and socio-economic inequalities (by promoting the health of the most well-off and relocation of jobs to high-income countries).
Spain	The academic discussion on ethical issues is rather scant. The main areas of focus in the Spanish academic debate are the military (autonomous weapon systems), work and medicine. With regard to robotics' effects on work, the potential for unemployment (especially for low-skilled workers) has been highlighted. There is some discussion on the impacts of driverless cars (discussed in a single article).
Sweden	The academic discussion on ethical issues with AI and robotics technologies is scant and very recent. All academic articles found were either students' master's or bachelor's theses, mostly in the areas of law and computer science. Ethical issues discussed in the articles relate to unemployment, worker safety, responsibility and liability, loss of control, privacy, intellectual property rights (for AI generated art), and "electronic personhood" for robots.
United Kingdom	The international and UK academic debate are so closely connected that they are hard to distinguish or detangle. It was hardly possible to identify a specifically "British" academic debate on ethics of AI and robotics. That being said, the few articles that were analysed mostly focused on issues in transportation, healthcare and robotics in general, involving values such as autonomy, safety, enablement, independence, responsibility, privacy and social connectedness.
United States	The academic debate on ethical issues in AI and robotics in the US has focused, first and foremost, on drones and autonomous robots, especially drone warfare, autonomous weapons and the ethics of their use and design. The second most common topic (though a distant second) was applications of AI and robotics in the medical industry. The third most common topic was the study of human acceptance of AI & robotics broadly (whether attitudes toward robots in the home or of AI in the doctor's office, as just a few examples). Ethical issues discussed included or related to: justice, equity, explainability, transparency, acceptance, autonomy, safety, accountability, liability, privacy, data protection, consumer confidence, regulation, certification, laws of war, and rules of engagement, amongst others.

Table 5: Summarised findings per country on the country-specific academic debate on ethical issues in robotics.

4.2.3. Popular media debate on ethical issues in AI

Country	Summarised findings
Brazil	There exist a lot of media articles that discuss the use of AI to improve education, and the importance of training people in the use of AI so as to better prepare them for the (future) job market. Media articles have more to say about the economic impact of AI development than the academic articles. They often talk about recent AI advancements and their economic impacts in general terms.

China	In China, the media articles that were found discuss the threat of AI development to human beings and society, provide analyses of the ethical issues, and discuss norms and regulation for AI technology. Ethical issues and values include safety, privacy, fairness, and control, and many articles have a wide scope. "Some authors suggest to carefully define the relationship between man and machine, and take advantage of Chinese traditional culture to establish human-machine relationship pattern." Some argue that ethics research needs to be reinforced, as well as the role of leading enterprises and "top-level design of AI". Suggestions are frequently offered to improve laws and regulation, industry norms and standards for AI, and to establish ethical values and principles for AI development. There is also recognition for the need to engage with stakeholders in AI development.
France	*An analysis of the popular media debate was not conducted.*
Germany	Much of the German popular media debate focuses on the same topics that are discussed in the academic debate, especially autonomous vehicles, AI and robotics in the workplace, and AI and robotics in the defence sector. The most important issues relating to these are: liability issues (for autonomous cars), automated decision-making by AI and robotic systems, the potential for mass unemployment and its societal effects, robots making decisions on life and death, and ethics by design ("Can we teach robots morality?").
Greece	The popular media discussion on ethical issues is burgeoning and very recent (mostly after 2018). There is discussion of potential impacts on the legal sector (potential for unjust rulings due to automated decision-making), democracy, and work and employment. Many texts also discuss the importance of designing and training AI systems to behave ethically.
Netherlands	A lot of the articles written about the implications of AI in the Netherlands specifically, are focussed on the social impact AI will have and the need for ethics in the development of AI applications. Not many articles specify the moral dilemmas of concepts that should be discussed. Rather, the take home message of most articles is that ethics is important.
Poland	Overall, there is little attention for ethical issues in relation to AI in Polish media, and especially for ethical issues in the Polish context. Articles focus mostly on very general question such as what may happen if AI systems become more intelligent than humans. Country-specific issues were raised with regard to AI's effects on jobs, AI's effects on the economy, algorithmic transparency, bias and discrimination, and privacy and data protection.
South Africa	There is a livelier ethical discussion in the popular media than in academia. Much of the debate is about autonomous weapons and their ethical issues, and about social justice (relocation of jobs to high-income countries, increasing unemployment in South Africa), and South Africa as a moral leader and what African values could bring to the regulation of AI and robotics.
Spain	Topics in the popular media debate are mostly similar to the academic discussion in Spain: ethical issues in relation to autonomous vehicles, autonomous weapons, autonomous decision-making, work/jobs, bias and discrimination, use of data, and privacy.
Sweden	The Swedish popular media debate analysis was based on six articles from online scientific and IT news magazines. In these articles, the following concerns in relation to AI and robotics were raised: governance of implementation of AI technology in Swedish society, workers' job security, distribution of welfare to future generations, longevity, cybersecurity and cyberwarfare, the human aspect of AI, and existential risks of AI.
United Kingdom	Media in the UK often point to the dangers that AI and/or robotics might bring about. Reported risks include, in particular, bias and discrimination, manipulation of opinions, and job losses. The only media analysis that was found notes on the media coverage of AI: "while we found some sensationalised content, we saw far less than expected" (p. 8.).
United States	In the US, popular media coverage of the ethical issues in AI runs the gamut of topics from hiring practices, education and replacing low-wage workers to military decision-making, facial recognition and immigration, next-generation finance and much more. The primary

	issues mentioned are displacement of human labour and bias and discrimination in all forms, most notably in facial recognition and hiring, as well as law enforcement and criminal justice. Science fictional "AI overlords" narratives are also quite common, but more recently, appear to be mentioned as inaccurate or overblown rather than as possible or feared futures.

Table 6: Summarised findings per country on the country-specific popular media debate on ethical issues in AI.

4.2.4. Popular media debate on ethical issues in robotics

Country	Summarised findings
Brazil	A lot of media articles discuss the use of robotics to improve education, and the importance of training people in the use of robotics so as to better prepare them for the (future) job market. Media articles have more to say about the economic impact of robot development than the academic articles. They often talk about recent robotics advancements and their economic impacts in general terms.
China	The is focus on ethical issues in relation to robotics in general and in relation to specific applications such as autonomous vehicles and "hotel service robots". Issues discussed include privacy, responsibility, job, and human control. Suggestions are frequently offered to improve laws and regulation, industry norms and standards for robotics, and to establish ethical values and principles for robotics development. There is also recognition for the need to engage with stakeholders in robotics development.
France	*An analysis of the popular media debate was not conducted.*
Germany	Much of the German popular media debate focuses on the same topics that are discussed in the academic debate, especially autonomous vehicles, AI and robotics in the workplace, and AI and robotics in the defence sector. The most important issues relating to these are: liability issues (for autonomous cars), automated decision-making by AI and robotic systems, the potential for mass unemployment and its societal effects, robots making decisions on life and death, and ethics by design (can we teach robots morality?).
Greece	The popular media discussion on ethical issues is burgeoning and very recent (mostly after 2018). There is discussion of potential impacts in the military domain (autonomous robots), and on work and employment. Many texts also discuss the importance of making sure robotic systems behave ethically.
Netherlands	There is a broad range of robots that are discussed in the popular media debate. Healthcare is a popular topic for robotics in the Netherlands, as there are already robots used in this field. Emerging robots like household / home robots, self-driving cars and police robots are often discussed as well. The main issues that are addressed on the short term or currently experienced are safety and privacy. On the long term, questions about responsibility and liability are mentioned.
Poland	Overall, there seems to be little attention for ethical issues in relation to AI in Polish media. The articles that have been analysed often consider robotics technology in general, while there was some focus on industrial robots, driverless vehicles and drones, and services and transportation applications. Ethical issues discussed related to the impacts in terms of unemployment and the quality of work, the impacts on international economic relations, safety, privacy, psychological implications of the interaction with humanoid robots, and criminal liability for damages.
South Africa	There is a livelier ethical discussion in the popular media than in academia. Much of the debate is about autonomous weapons and their ethical issues, and about social justice (relocation of jobs to high-income countries, increasing unemployment in South Africa), and South Africa as a moral leader and what African values could bring to the regulation of AI and robotics.

Spain	Topics in the popular media debate are mostly similar to the academic discussion in Spain: ethical issues in relation to autonomous vehicles, autonomous weapons (killer robots), work/jobs, human-robot interaction, and privacy.
Sweden	The Swedish popular media debate analysis was based on six articles from online scientific and IT news magazines. In these articles, the following concerns in relation to AI and robotics were raised: governance of implementation of AI technology in Swedish society, workers' job security, distribution of welfare to future generations, longevity, cybersecurity and cyberwarfare, the human aspect of AI, and existential risks of AI.
United Kingdom	Media in the UK often point to the dangers that AI and/or robotics might bring about. Reported risks include, in particular, bias and discrimination, manipulation of opinions, and job losses.
United States	In the US, there appears to be less coverage in general of ethical issues in robotics, simply in terms of numbers of articles, which may be because "AI" serves as a more useful umbrella term for hybrid technologies such as driverless cars. Discussion of military autonomous weapons and the ethics of their use dominates popular media coverage, perhaps because it is so potentially sensational. Driverless cars are also a prominent feature of the media landscape. After that, robot companions, whether recreational or medicinal, are probably the next most common topic. The effect of ubiquitous robots on the workforce, typically in manufacturing, driving and the service and medical industries, is the most common social issue discussed.

Table 7: Summarised findings per country on the country-specific popular media debate on ethical issues in robotics.

4.3. Discussion of findings

Upon preparing the SIENNA country studies task for the ethical analysis of AI and robotics, we hoped the results would lead to the identification of new ethical issues not found in the broader literature. Unfortunately, however, few unique insights about ethical issues were gleaned. That said, we can still draw a number of interesting conclusions about the findings laid out in the previous subsection, and highlights some of the similarities and differences between the debates in the twelve countries under study. Please note that this section does not qualify as a proper comparative analysis of the findings since, due to time constraints, we have not been able to follow the rigorous standards required for such an analysis.

Regarding the academic debates in the twelve countries, the following has been observed. In some countries, there seems to be a preponderance of articles on a broad range of topics that are representative of a national academic debate on ethics AI and robotics in those countries (i.e., in China, Germany, the United States), whereas in other countries the national academic debate has been more modest (i.e., in France), and in still others it has been rather scant (i.e., in Brazil, Greece, Poland, South Africa, Spain, Sweden, United Kingdom). Especially notable is the relative lack of country-specific studies in the UK, which may be explained by the international academic orientation of UK institutions.

Across all twelve countries, the most widely discussed application areas of AI and robotics are defence, medicine, transportation, and the workplace, with the autonomous weapon systems (especially "killer robots"), care robots, healthcare apps, surgical robots, sex robots, and autonomous vehicles being the most-discussed products. Especially notable was the significant amount of attention the ethics of defence applications of AI and robotics has been receiving in most countries. Overall, a wide range of ethical issues have been discussed—which largely seems to be reflective of the wider international debate—that includes those relating to justice, equality, autonomy, dignity, explainability, transparency, safety, accountability, liability, privacy, and data protection. Of these, perhaps the most frequently mentioned issues concern justice, privacy, and safety, which were often still addressed in countries were academic discussion was found to be scant.

The national academic debates in the US, Germany and China stood out from the rest in that they also focused on potential broad-scoped solutions to the ethical issues with AI and robotics, including through laws, standards, and regulation, as well as through ethics by design (training systems how to behave ethically) and investigating whether moral reasoning systems can potentially be implemented in robots and AI systems.

Regarding the popular media debates within the twelve countries under study, the following has been observed. In all of the countries, with the possible exception of Poland, there has been substantial debate in the national popular media on ethical issues in relation to AI and robotics—although in some countries the debate has only recently become more intense. Often, it was found that the application areas, products, and ethical issues and principles addressed in the popular academic debate largely mirror those in the academic debate. Issues related to the potential economic effects of AI and robotics technology, however, seem to get slightly more attention. Also, in what can be regarded as a somewhat curious finding, there seems to be less (country-specific) discussion of issues to do with sex robots in the national popular media than in academia in most countries.

5. Ethical analysis: General ethical issues in AI and robotics

In this section, we identify and analyse the main ethical issues with regard to artificial intelligence and robotics technology *at large*. In the methodology section of this deliverable, we have indicated that in conducting our ethical analyses, we follow the *Anticipatory Technology Ethics* approach developed by Brey (2012).[36] This means that the ethical issues in relation to AI and robotics will be analysed at three so-called *levels of ethical analysis*: (1) the *technology level*, the most general level of description, which specifies the technology in general, its subfields, and its fundamental techniques, methods and approaches; (2) the *artefact level* or *product level*, which provides a systematic description of the technological artefacts (physical entities) and procedures (for achieving practical aims) that are being developed on the basis of the technology; and (3) the *application level*, which defines particular uses of these artefacts and procedures in particular contexts by particular users.

The present section covers our ethical analysis at the first of these three levels, namely, the technology level. Our objects of analysis at this level consist of the aims of the fields of AI and robotics and their subfields, the fundamental techniques, methods and approaches used in these fields, and the general implications and risks resulting from artefacts and applications of the fields. For instance, in this section, we discuss the ethical issues in relation to such general aims of AI and robotics as the *improvement of efficiency and productivity*, and the *reduction risks*. Also, we discuss the ethical issues inherent in such AI and robotics techniques, approaches and concepts as *machine learning*, *algorithms* and *robot sensing*, and we detail the ethical issues with respect to, for example, *mass unemployment*, *justice and fairness*, and *safety and security*.

In this section, we focus on both present issues and issues that may occur between now and 20 years into the future. This section therefore draws on SIENNA *foresight* analyses (mainly through expert workshops and expert interviews) that have been conducted to (1) obtain descriptions of the possible, plausible or probable future development of AI and robotics technologies, their products, and their applications, as well as to (2) identify potential ethical issues in relation to these technologies, products and applications. Most of our input for this section, however, consists of an extensive analysis of the academic and popular literature on general ethical issues in AI and robotics. In addition, we have on occasion used ethical checklists to perform our own analysis in areas where the literature was scarce.

This section is structured as follows. Subsections 5.1 and 5.2 describe the general ethical issues in AI technology and the general ethical issues robotics technology, respectively. In turn, each of these subsections consists three subsections that detail ethical issues with regard to (1) the general aims of the fields and its subfields, (2) their techniques, methods and approaches, and (3) their general implications and risks.

5.1. General ethical issues in AI

This subsection offers a discussion of the general ethical issues in artificial intelligence (AI). We begin, in subsection 5.1.1, by describing the ethical issues that are inherent in the general aims of AI and its subfields. Then, in subsection 5.1.2, we detail for the most important AI techniques, methods and

[36] Brey, 2012, op. cit.

approaches, the main ethical issues that are specific to them (i.e., issues that are inherent in, or frequently occur with, these techniques, methods and approaches). Finally, in subsection 5.1.3, we describe the main ethical issues with regard to some of the general implications and risks of AI technology (e.g., harms to autonomy, privacy, justice). Figure 2 offers an overview of the structure of these three subsections.

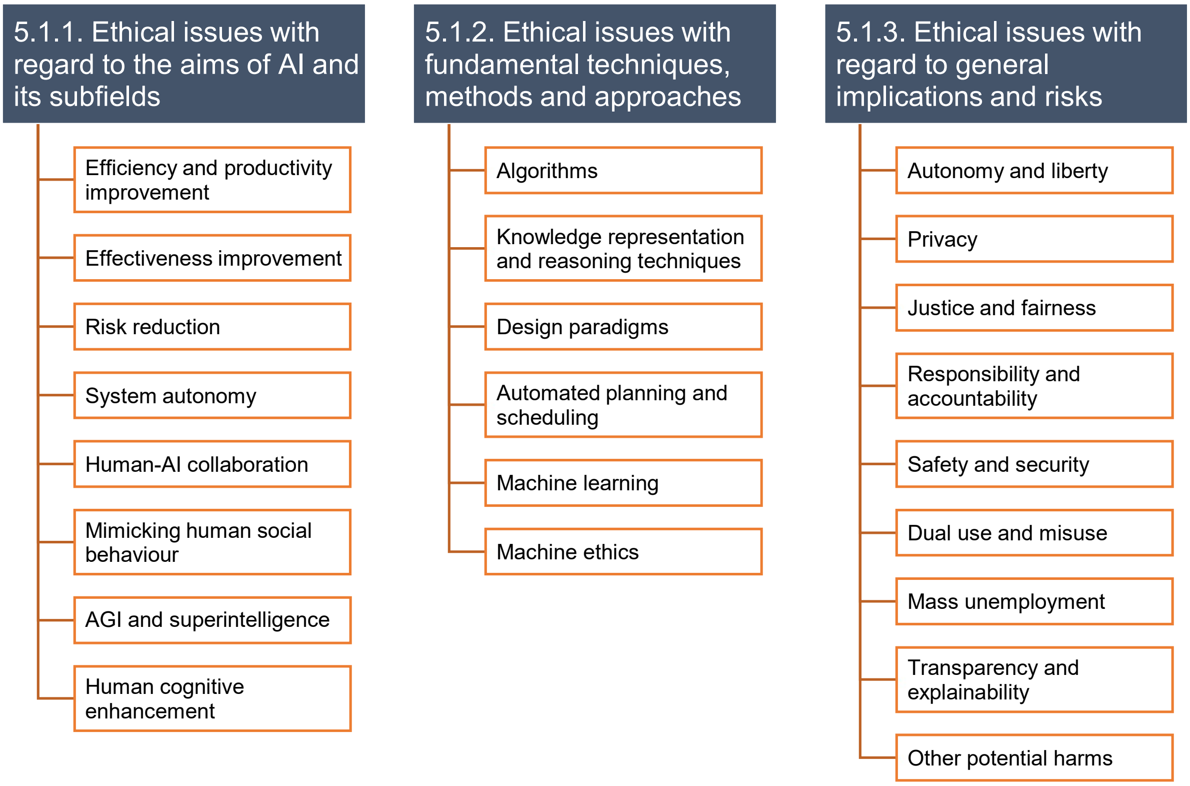

5.1.1. Ethical issues with regard to the aims of AI and its subfields
Efficiency and productivity improvement
Effectiveness improvement
Risk reduction
System autonomy
Human-AI collaboration
Mimicking human social behaviour
AGI and superintelligence
Human cognitive enhancement
5.1.2. Ethical issues with fundamental techniques, methods and approaches
Algorithms
Knowledge representation and reasoning techniques
Design paradigms
Automated planning and scheduling
Machine learning
Machine ethics
5.1.3. Ethical issues with regard to general implications and risks
Autonomy and liberty
Privacy
Justice and fairness
Responsibility and accountability
Safety and security
Dual use and misuse
Mass unemployment
Transparency and explainability
Other potential harms

Figure 3: Structure of subsection 5.1 on general ethical issues in artificial intelligence.

5.1.1. Ethical issues with regard to the aims of AI and its subfields

In this subsection, we identify and analyse the ethical issues associated with the most important aims and sub-aims in the development of AI systems. In SIENNA Deliverable 4.1,[37] we have stated that the primary aims of AI research are, firstly, to systematically study the phenomenon of intelligence, and secondly, to develop programs and tools that can automate intelligent behaviour such as information gathering, detecting, planning, learning, communicating, and manipulating. Since the second aim is most relevant for the study ethical issues in relation to AI (as it lies at the basis of real-world applications of AI), we largely focus on this aim and break it down into various sub-aims.

We have identified the following ethically relevant aims and sub-aims of AI: *efficiency and productivity improvement*; *effectiveness improvement*; *risk reduction*; *system autonomy*; *human-AI collaboration*; *mimicking human social behaviour*; *artificial general intelligence and superintelligence*; and *human cognitive enhancement*. For each of these, we discuss below the most important ethical issues.

Efficiency and productivity improvement

One of the main drivers in the development of applied AI technology is the expectation that its use will result in significant improvements in efficiency and productivity. In many domains, AI technology is

[37] Jansen, et al., 2018, op. cit.

being developed on the assumption that it that helps achieve more at lower costs in terms of expended time, money, effort and/or risk. The monetary cost reductions that are sought are usually labour cost reductions. While the objective of productivity and efficiency improvements in various sectors promises economic benefits for organisations and society at large (and may perhaps improve worker well-being through a reduction in the repetitiveness of human tasks), it may also bring with it inherent risks in terms of job losses, especially in routine and low-skill labour, and in terms of un-fulfilled demands for highly skilled workers (to service the more complicated systems that AI-based efficiency improvement may require). The former may give rise to issues of inequality as a result of rising unemployment among particular groups in society, and the latter may result in undertraining of workers and associated risks of workplace and societal harms caused by AI system failures.

Overall job losses as a result of automation technology between 2018 and the mid-2030s have been estimated at around 30%.[38] It should be noted, however, that AI-induced job displacement is expected to be offset to some extent by rising real income levels as a result of higher productivity and lower prices for products, which would allow for increased consumer spending and higher job creation.[39] The added jobs are likely to require either highly "human" (creative and social) skills or highly technical skills,[40] and the extent to which they will be able to compensate for job losses depends on how big the demand for them will be. This means that workers trained for routine and low-skill (technical) work may face an uncertain future, and that a shortage of highly skilled technical workers to design and maintain AI systems may develop. (More on the ethical issues surrounding the potential for mass unemployment in subsection 5.1.3.)

Effectiveness improvement

Another aim in the development of AI technology, which is closely related to the aim of efficiency and productivity improvement, is the aim of making systems that are more effective than humans in particular tasks. In many tasks, AI systems now match or exceed human-level performance in terms of quality of the results (e.g., in certain visual categorization tasks), and there are many things which AI systems can now do that have not previously been possible (e.g., being driven in an autonomous car).

The aim of effectiveness improvement may hurt workers in ways similar to those of efficiency and productivity improvement: job losses that may or may not be sufficiently compensated for by the creation of new jobs, with workers in routine, low-skill jobs being most vulnerable, and un-fulfilled needs for highly skilled workers. Completely new innovations that are designed to serve previously unknown needs (and are therefore not replacing existing practices) may have less of an impact in terms of job losses.

Risk reduction

A further aim in the development and implementation of AI technology is reducing risks to humans in a variety of applications. Risk reduction is an aim in many areas where AI systems are more efficient and effective at performing certain tasks than humans are. Such areas include medicine (e.g., AI

[38] Hawksworth, John, Euan Cameron, and Richard Berriman, "Will Robots Really Steal Our Jobs?: An International Analysis of the Potential Long-Term Impact of Automation," *PricewaterhouseCoopers*, 2018. https://www.pwc.co.uk/economic-services/assets/international-impact-of-automation-feb-2018.pdf.

[39] Hawksworth, John, and Yuval Fertig, "AI and robots should create as many jobs as they displace in the long run," *PricewaterhouseCoopers*, 2018. https://pwc.blogs.com/economics_in_business/2018/07/ai-and-robots-should-create-as-many-jobs-as-they-displace-in-the-long-run.html

[40] Ibid.

systems detecting heart arrhythmias on the basis of electrocardiograms[41]), transportation (e.g., intelligent camera systems designed to catch drivers who are using their mobile phones[42]) defence (e.g., AI-assisted tablets for soldiers to improve communication and awareness[43]), and others.

Ethical concerns with regard to the aim of risk reduction through AI systems may relate to safety and equality. Risk reduction may induce a false sense of security if the capabilities and workings of the AI systems are not well understood, and may lead to complacency (e.g., in the medical domain) or overconfidence (e.g., in the military domain). Medical professionals may grow overly reliant on AI systems for medical diagnosis and be tempted to put less effort in performing independent assessments themselves, thus placing patient's safety at risk (assuming the AI systems will never provide completely infallible results). Similarly, soldiers may overestimate the degree to which AI systems are providing them with better protection and offensive capabilities, and may therefore take undue risks to their and other people's safety. Finally, where risk reduction in AI systems is targeted at specific (groups) of individuals, it may put other (groups of) people at an unfair relative disadvantage.

System autonomy

An important sub-aim of the aforementioned aim of efficiency and productivity improvement is enhancing the autonomy of technical systems. To improve cost-efficiency often means to lower the costs of relatively expensive human labour. And to enhance productivity often means to implement faster or larger-scale production processes. To achieve both, it often helps to make use of AI-enabled systems with high levels of autonomy.

It is mainly through the desire for highly autonomous behaviour by AI systems that the increased use of such systems raises the spectre of unemployment in certain sectors of the economy and un-fulfilled demands for highly skilled workers. Human workers cannot compete with autonomous AI systems that can do the same work in more cost-efficient and faster ways. A drive toward autonomous systems may also raise issues in terms of a general deskilling among the population at large (or at least among individuals who are not tasked with servicing the AI systems). People may unlearn many of the skills they needed to perform certain tasks before AI systems took them on. The erosion of such skills in individuals may put their and other people's safety and well-being at risk when the AI-based systems are out of order.

Some wholly different issues relating to the aim of AI autonomy are issues relating to accountability and responsibility for the behaviour of autonomous AI systems. There are as yet no clear answers as to who is responsible for the proper functioning of autonomous AI systems (e.g., designers, owners, users) and who should be held accountable or liable in the event something goes wrong.[44] These issues have been made more complicated by the development AI systems whose internal workings are not transparent (e.g., systems based on neural networks), and will be critically important in applications

[41] Rajpurkar, Pranav, Awni Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Ng, "Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks," *Cornell University arXiv*, 2017. https://arxiv.org/pdf/1707.01836.pdf

[42] Spilka, Dmytro, "How AI Is Keeping Us Safe From Drivers Who Use Their Mobile Phones at the Wheel," *Datafloq*, February 15, 2019. https://datafloq.com/read/aikeeping-safe-drivers-using-mobile-phones-wheel/6064.

[43] Patterson, Dan, "How AI-Powered Robots Will Protect the Networked Soldier," *TechRepublic*, April 6, 2016. https://www.techrepublic.com/article/how-ai-powered-robots-will-protect-the-networked-soldier.

[44] Matthias, Andreas, "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata," *Ethics and Information* Technology, Vol. 6, No. 3, 2004, pp. 175–183.

that can involve life-and-death situations. (More on the ethical issues surrounding responsibility and accountability in subsection 5.1.3.)

Finally, the creation of highly autonomous AI systems may have implications for interpersonal relations among humans. If people rely too much on such systems, their contact with other humans in day-to-day activities (e.g., dealing with sales clerks, personal contact at work) might decrease as a result, thus potentially affecting their well-being. In certain situations (e.g., in healthcare), there may even be harms to human dignity.

Human-AI collaboration

The aim of human-AI collaboration stands somewhat in opposition to the aforementioned aim of enhancing system autonomy. Recent research has found that in many applications, AI can perform some tasks better than human, but never all of the tasks that are part of those jobs.[45] Humans and AI systems can complement each other: humans generally possess stronger leadership, teamwork, creativity, and social skills, whereas AI systems have better speed, scalability, and quantitative capabilities.[46] To maximise efficiency and productivity, AI systems will therefore need to be designed to collaborate with humans in an efficient manner (e.g., in the form of decision-support systems or collaborative robots). Furthermore, the creation of partnerships between humans and AI systems may be incentivised since the safety and continuity of production processes may benefit from having humans in the loop, and because such partnerships ensure employment opportunities for humans.

The aim of human-AI collaboration raises a number of potential ethical issues. Firstly, working in close proximity to, and collaboration with, an AI-based system, whose behaviour may not be perfectly predictable, may increase safety risks for the human collaborator(s) and others, especially in industrial settings. Secondly, when humans are taking cues from AI systems, they may have a right to explanations of how these systems arrive at particular decisions, which for important kinds of (neural-network-based) AI systems are difficult to provide. Thirdly, in human-AI collaboration, there is a risk of humans unduly influencing AI systems by, for example, feeding them with biased data, which may lead to bad decisions. Further issues may include the risk of deskilling (i.e., humans not knowing how to complete the task by themselves), the potential for human-AI interaction to reduce human-to-human social interactions, and possible negative effects on privacy (e.g., AI systems monitoring their human collaborators). In light of these issues, there may be a strong argument for explicitly embedding ethical principles and into collaborative AI systems,[47] which itself may also present a number of difficult challenges.

Mimicking human (and animal) social behaviour

Another aim in the development of certain types of AI systems is to mimic the capacities of humans and animals for social behaviour, which would enable these systems to interact with humans in socially intelligent ways. Capabilities such as engaging in natural conversation with humans and understanding human emotions are highly coveted. The development of socially intelligent AI systems, robotic and

[45] Brynjolfsson, Erik, and Tom Mitchell, “What Can Machine Learning Do? Workforce Implications,” *Science*, American Association for the Advancement of Science, December 22, 2017. http://science.sciencemag.org/content/358/6370/1530.

[46] Daugherty, Paul R., and H. James Wilson, “How Humans and AI Are Working Together in 1,500 Companies,” *Harvard Business Review*, April 4, 2019. https://hbr.org/2018/07/collaborative-intelligence-humans-and-ai-are-joining-forces.

[47] Rossi, Francesca, “Human-AI Collaboration: Technical & Ethical Challenges,” *OECD Conference*, October 26, 2017. http://www.oecd.org/going-digital/ai-intelligent-machines-smart-policies/conference-agenda/ai-intelligent-machines-smart-policies-rossi.pdf

non-robotic, can benefit applications in such domains as healthcare, education, retail and entertainment.

Ethical issues related to this aim may include, firstly, the potential for decreased human social interaction among individuals who deal with socially intelligent AI systems, given that interaction with such systems may compete with, and in part replace, human interaction. Diminished human-to-human social interaction resulting from the use of AI systems may harm individuals' wellbeing on the assumption that, at least for the foreseeable future, AI systems will not come close to being able to perfectly emulate the full breadth and depth of human social intelligence and communicative abilities. In addition, reduced human social interaction may result in a general social deskilling among individuals, as they are less exposed to opportunities to hone their human-to-human social skills. Secondly, the substitution of humans by socially intelligent AI systems may, in applications where social interactions have critical functional importance, result in poorer task outcomes, thereby further harming human wellbeing. Thirdly, the aim of mimicking human social behaviour may raise issues of trust and deception, as humans may be tricked into believing the socially intelligent AI system represents a real person.

Artificial general intelligence and superintelligence

A present aim of more fundamental research in AI is the development of artificial general intelligence (AGI), and a future aim may be to develop artificial superintelligence. Even if many experts indicate that we are not likely to see either of these being realized within the next 20 years (which is SIENNA's scope for studying the ethical issues in AI and robotics),[48,49] it may still be worthwhile to briefly consider the ethical issues related to these aims.

Firstly, AGIs may (or may not) be developed that possess consciousness like humans (thus fitting John Searle's definition of "strong AI"[50]), and they may experience suffering as a result, especially if it turns out humans are unwilling or unable to accord them certain legal rights. Secondly, AGIs might make humans completely obsolete in many economic sectors, which may have major consequences for human wellbeing and equality, amongst other values. Thirdly, there is the potential that AGIs and superintelligences may have aims and values built into them that are badly designed, with very negative (intended or unintended) consequences for humans (e.g., a superintelligence that seeks to manufacture as many paperclips as possible and is willing to kill humans should they be an obstacle to reaching this goal[51]). Fourthly, the quest to build an AGI may eventually initiate a runaway reaction of self-improvement cycles by AGIs (i.e., an "intelligence explosion"), culminating in what is called a "technological singularity". This may have very profound and difficult-to-predict consequences for humans and society.

Human cognitive enhancement

A final aim in the development of important kinds of AI systems is to enhance human mental capabilities, and to treat or compensate for neurological damage in humans. This aim may become

[48] SIENNA interviews with AI experts (N=5) held in January, 2019.

[49] In a more extensive survey of AI experts, "[t]he median estimate of respondents was for a one in two chance that high-level machine intelligence will be developed around 2040-2050, rising to a nine in ten chance by 2075. Experts expect that systems will move on to superintelligence in less than 30 years thereafter." Müller, Vincent C., and Nick Bostrom, "Future Progress in Artificial Intelligence: A Survey of Expert Opinion," Fundamental Issues of Artificial Intelligence, 2016, pp. 555–572.

[50] Searle, John R., *Mind, Language and Society: Philosophy in the Real World*, Phoenix, New York, 1999.

[51] Bostrom, Nick, "Ethical Issues in Advanced Artificial Intelligence," 2003. https://nickbostrom.com/ethics/ai.html.

more prominent in the future than it is currently. For example, AI may be used in brain-computer interfaces to augment human intelligence and in neural prostheses to replace a missing or damaged neurological functionality.

Ethical issues related to the aim of human cognitive enhancement may be severe. Firstly, cognitive enhancement through AI technology may negatively impact social equality, since not all humans may have access to such cognitive enhancement. Given that one's intelligence is arguably one of the fundamental factors to success in life, the impact on equality can be profound. Secondly, cognitive enhancement can have severe harmful impacts on human psychology and identity, and by extension human dignity. Finally, there may be harmful effects on privacy as a result of AI-based cognitive enhancement technology, as it may become possible to eavesdrop on neural prostheses and computers that interface with the brain. (For a more detailed analysis of the ethical issues surrounding the potential for human enhancement through AI technology, please see SIENNA Deliverable 3.4: Ethical Analysis of Human Enhancement Technologies.)

5.1.2. Ethical issues with regard to fundamental techniques, methods and approaches

In this subsection, we describe for the most important fundamental techniques, methods and approaches in AI the main ethical issues that are specific to them (i.e., issues that are inherent in, or frequently occur with, these techniques, methods and approaches). Please note that our listing of AI techniques, methods and approaches is not exhaustive; we have attempted to identify only those techniques, methods and approaches that may give rise to significant and specific ethical issues.

Algorithms

An algorithm is a sequence of instructions that in specifies an unambiguous manner how to solve a class of problems or perform a certain task. Algorithms do not only exist in computing; they also exist in mathematics, and are implemented in biological neural networks and electrical circuits. Computer algorithms are algorithms that are implemented in a formal programming language and are part of a computer program. A computer program centrally consists of algorithms and can even itself be considered to be a complex algorithm. Algorithms are effective methods for producing a result. They start from an initial state with (optional) initial input, and then describe a computation that involves a finite number of well-defined successive states that results in eventual "output" and a final ending state. The instructions from going from state to state can be described as rules. For example, an algorithm can contain a rule specifying that if the input consists of the letter "y", then display the text "Are you sure?" on the screen and wait for further input.

At first glance, it might be believed that although algorithms may be used in programs that raise moral issues (for example, programs designed to collect personal information without consent, or programs that can copy themselves and infect a computer), algorithms themselves are morally neutral. Take, for example, an algorithm that calculates the sum of two numbers: what could possibly be morally controversial about it? Similarly, an algorithm within a car navigation system that calculates the shortest route between two points does not seem to raise any moral issues. So, can there be an ethics of algorithms?

There is an emerging consensus that many algorithms are not morally neutral because they are value-laden: they have orientations in favour of or against certain values.[52,53] As Kraemer, van Overveld and Peterson argue,[54] they can be conceived of an instance of a broader phenomenon, which is that technological artefacts can be value-laden (see also: Van den Hoven, Vermaas and van de Poel, 2015;[55] Brey, 2010[56]). These authors are not making the claim that all algorithms are value-laden. Presumably, an algorithm that merely adds up two numbers is not value-laden in any interesting sense. However, as Kraemer et al. claim, many algorithms are value-laden in that they cannot be designed without implicitly or explicitly taking a stand on ethical issues. There are multiple ways of designing them to perform the specified task, and different designs involve different value choices.

It is often possible to design different algorithms to perform the same task. For example, a chess program can employ different algorithms to play chess, for example ones that do exhaustive searches of several moves ahead, or ones that instead focus on investigating a limited set of moves. Different algorithms can exist at the algorithmic (logical) level for the same task, and they can then also be implemented differently in programming. Moreover, specified tasks that algorithms need to carry out are often not defined in a formal manner, but are defined using terms and concepts from ordinary language that includes vagueness, ambiguities, and unstated background assumptions. For example, an algorithm that is to identify running behaviour in a video feed must translate a vague concept, “running”, into an exact specification, and there are multiple ways to do that. In addition, there are often additional requirements, explicitly stated or implicit, that algorithms must satisfy that could affect its design. For example, a navigation algorithm may be designed to calculate the shortest distance between two points, but requirements may be added that waterways and unpaved roads are excluded, or that the vehicle does not cross borders.

So, algorithm design often involves choice. The next argument to make is that some of these choices are morally charged. That this is sometimes so can be seen by considering two central functions that algorithms have. Some algorithms have an informational function: the outcome they generate is a piece of information (a number, a string, a record, a picture) that can then be used by either humans or machines. (They can also be input for other algorithms.) Other algorithms rather have the function to recommend or cause action: they issue a particular recommendation to human users (or machines), as when a navigation system tells a driver to make a left turn, or they cause certain events to happen, as when an algorithm embedded in a robot causes it to raise its arm.

It is easiest to see for those algorithms that recommend or cause actions that they can be morally charged. Actions, in general, may be moral or immoral, so it follows that if an algorithm recommends or causes an action, it takes a moral position. Not all actions involve significant moral choices, of course, but a good many of them do. So, for example, algorithms that recommend or cause actions that violate people’s rights or are discriminatory are clearly not morally neutral.

It can moreover be shown that moral choice is often involved in algorithms that do not recommend or cause actions but merely produce information. The production of information is a process that involves

[52] Kraemer, Felicitas, Kees van Overveld, and Martin Peterson, “Is there an ethics of algorithms?,” *Ethics and Information Technology*, Volume 13, Issue 3, 2011, pp. 251–260.

[53] Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S. and Floridi, L. (2016). The Ethics of Algorithms: Mapping the Debate.

[54] Kraemer, et al., 2011, op. cit.

[55] Van den Hoven, J., Vermaas, P. & Van de Poel, I. (Eds.), *Handbook of Ethics, Values, and Technological Design. Sources, Theory, Values and Application Domains*. Dordrecht: Springer.

[56] Brey, P. (2010). Values in Technology and Disclosive Computer Ethics. In L. Floridi (Ed.), *The Cambridge Handbook of Information and Computer Ethics* (pp. 41-58). Cambridge: Cambridge University Press.

the selection and interpretation of data, and the use of standards of evidence for drawing conclusions from data, and the use of categories to interpret and categorize data. All of these processes can be construed as actions that involve choice, and in some cases these choices can be seen to be morally charged.

To begin, the use of certain categories to represent reality involves moral choices. Some categories, for example, are morally controversial by grouping or depicting entities in a way that some say they should not be grouped or depicted. It would, for example, be morally controversial to have an algorithm that classifies people as “racially pure” and “racially impure”. Similarly, it involves an (often implicit) moral choice to employ only two categories for categorizing gender (“male” and “female”), thereby excluding the existence of non-binary genders. In general, the choice of categories used in algorithms and in the representation, interpretation, categorization and organization of data, involves implicit or explicit choices to highlight or “construct” certain aspects of reality, while downplaying or leaving out other aspects, and to invoke certain attitudes in users and prime them in a certain way.[57] Some of these choices are moral in nature.

The inferences drawn by algorithms can also be morally charged. Except for logically valid inferences, inferences tend to be underdetermined by the evidence. Algorithms may, for example, make generalizations based on a limited number of positive instances, or assume causal relations where there are only statistical correlations. Such inferences are not always morally charged. For example, the inferences drawn by an algorithm from data from a quantum physics experiment are not likely to involve implicit moral choices. In other cases, however, inferences may be based on moral biases or prejudices. For example, algorithms may be structured to make prejudicial inferences to associate low socioeconomic status with crime. When no prejudices are involved, algorithms may also involve implicit moral choices. Kraemer et al. give the example of MR-scans of the heart, in which the algorithms that produce the image contain a threshold value for categorizing parts of an image as light or dark grey. This threshold value influences whether an MR-scan is classified as indicating possible pathology, and can create a bias towards false positives or false negatives.[58] But whether there are more false positives or false negatives is an implicit moral choice: it is a choice between avoiding inconvenience to a lot of people and unnecessary tests and avoiding undetected pathologies.

We have seen that algorithms can be morally charged for two broad reasons: either because the actions that they take or recommend involve moral choices, or because the inferences they draw and categories they use involve moral choices. Orthogonal to these two types of value-ladenness is the notion of *algorithmic bias*. Algorithmic bias is a type of value-ladenness of algorithms that results in unfair outcomes, either disadvantaging social groups (gender, race, ethnicity, age, etc.), people with certain characteristics (e.g., people whose surname is more than ten integers long, people with dual citizenship) or randomly selected individuals or groups. It can be found in categories used, inferences drawn, decisions made and actions taken. It may also result from a bias in the data used (see the part on “Justice and fairness” in subsection 5.1.3).

A third general way in which algorithms can be value-laden is by the degree to which they can be understood by their users and stakeholders. This specifically relates to algorithms that make decisions or recommend choices. *Algorithmic transparency* is the principle that the purpose, inputs and operations of algorithms must be knowable to its stakeholders. Advocates, such as the High-Level Expert Group on AI of the European Commission, hold this to be a moral principle: those affected by

[57] Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind.* University of Chicago Press.
[58] Kraemer, et al., 2011, op. cit.

an algorithm should have the ability to understand why the algorithm makes the decisions that it makes.[59] This is considered especially important in cases in which the rights of people are affected by the algorithm's decisions, for example in cases in which computer programs provide sentencing guidelines or decide on the creditworthiness of loan applicants. Algorithmic transparency is also considered to be a requirement for algorithmic accountability, which is the principle that organizations that use algorithms should assume responsibility for the decisions made by those algorithms.[60,61]

Knowledge representation and reasoning techniques

Knowledge representation is a subfield of AI that concerns itself with the fundamental challenges faced in representing information about the world in a form that a computer system can utilize to solve complex tasks (e.g., diagnosing a medical condition or having a dialog in a natural language). Things that an AI system may need to represent include: concepts, categories, objects, properties, situations, events, states, time, and causes and effects. Techniques (or rather, languages) for representing knowledge include first-order logic, modal logic, description logic, rules, frames, and semantic networks, amongst others.[62] Knowledge representation goes hand in hand with the capacity for automated reasoning about that knowledge. Techniques for automated reasoning include classical logics, fuzzy logic, Bayesian inference, reasoning with maximal entropy, and a large number of less formal ad-hoc techniques.[63] Knowledge representation and automated reasoning are foundational to the development of expert systems (see subsection 6.1.2 on knowledge-based systems).

As of now, there seems to be very little literature that directly addresses any ethical issues inherent in knowledge representation and reasoning techniques. However, drawing from our own limited ethical analysis, we can nonetheless offer a brief listing of potential ethical issues with regard to this subfield. First of all, to the extent that the task of representing knowledge in a knowledge base involves human interpretation and representation of facts,[64] there may be a risk of misrepresentation of that knowledge. Knowledge engineers may misrepresent or omit facts or concepts knowingly and potentially with malicious intent, or unknowingly, perhaps as a result of subconscious bias. Misrepresented knowledge in a knowledge base can have grave consequences depending on the purpose of the knowledge base. An example can be the inaccurate diagnosis of diseases in minorities resulting from a medical expert system's knowledge base that lacks well-established information on how the manifestation of disease symptoms differs among racial groups. Furthermore, any influential knowledge base in which knowledge is misrepresented or omitted as a result of cultural bias may do significant damage to cultures if it is exported to other parts of the world.

Related to these issues raised by hand-crafting ontologies are issues that may result from a necessary trade-off between expressivity or comprehensiveness of the knowledge representation, and inferential efficiency. It is generally held to be a practical impossibility to specify *all* preconditions—

[59] High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*, European Commission, 2019. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

[60] Binns, R. (2017). Algorithmic Accountability and Public Reason. *Philosophy and Technology*, 1-14.

[61] Mittelstadt et al., 2016, op cit.

[62] Trentelman, Kerry, *Survey of Knowledge Representation and Reasoning Systems*, Defence Science and Technology Organisation, Edinburgh, S. Aust., 2009. https://apps.dtic.mil/dtic/tr/fulltext/u2/a508761.pdf

[63] Ibid.

[64] For example, the Cyc project—arguably the world's longest-lived artificial intelligence project—has attempted to build a comprehensive knowledge base composed of common-sense rules (and assertions based on these rules) was largely created through hand axiom-writing, and as of 2017 has involved more than 1,000 person-years of effort in it construction. https://www.cyc.com/

including many common-sense ones—that will guarantee a particular action's successful execution.[65] Similarly, it is practically impossible to specify all of the effects an action might have.[66] This means that, for the sake of inferential efficiency in knowledge-based systems, inferential accuracy will have to be sacrificed to some degree. Even the most complex systems built using comprehensive knowledge bases and state-of-the-art inference engines cannot be trusted to provide 100 percent perfect results all of the time. In situations where trust in the veracity of the output of well-built knowledge-based systems is very high and where such systems are meant to replace human experts, this may potentially lead to various risks and harms as a result of overconfidence in the system's performance and insufficient oversight.

A potential third set of issues is raised by the prospect that knowledge representation in the future will evolve under machine control. It has been argued that in order for knowledge-based AI systems to cope with complex and ever-changing real-world environments, it will not be sufficient to simply add some facts or rules to their knowledge bases and reasoning systems.[67] Rather, the *way* these systems represent facts and concepts (i.e., their representational language, its syntax and semantics) must be automatically changeable.[68] Some authors hold that automatic representation development, evolution, and repair by AI systems should be an important objective in AI research in the next 50 years.[69] Such a development would raise issues of responsibility and accountability since designers and users would not be able to foresee unintended ethical effects of the automatic changes in AI system's representational language.

Automated planning and scheduling

Automated planning and scheduling (also known as AI planning) is concerned with the planning of specific tasks in order to achieve a pre-stated goal by going through a process of evaluating the outcomes of its potential courses actions and selecting the most favourable ones. AI planning is concerned with the computational study of this process. Planning is a capacity that is commonly associated with intelligent beings. Implementing planning behaviour in artificial agents may help us to understand intelligence. Planning tasks broadly consists of three components: a state-transition system (that represents the to-be-affected situation), a planner (that regulates the plans and policies used to reach set objectives), and a controller (that reacts to the planner's output).[70] If the system is time dependent, the planner commonly includes a scheduler as well. The scheduler is meant to solve time specific issues, such as starting one action, but not finishing it due to a dependence on another action that first needs to be completed.[71] These three parts together enable the algorithm to change its states and perform actions in order to reach a desired goal.

Let us discuss now some of the ethical issues with automated planning and scheduling. First of all, automated planning and scheduling techniques potentially raise safety issues, as they can make users

[65] In philosophy and AI, and knowledge-based systems in particular, this is referred to as the *qualification problem*. Reiter, Raymond, *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems*, Cambridge, Massachusetts: The MIT Press, 2001.

[66] This is what is known as the *ramifications problem* or *frame problem*. Ibid.

[67] Bundy, Alan, and Fiona Mcneill, "Representation as a Fluent: An AI Challenge for the Next Half Century," *IEEE Intelligent Systems*, Vol. 21, No. 3, 2006, pp. 85–87. https://ieeexplore.ieee.org/abstract/document/1637360

[68] Ibid.

[69] Ibid.

[70] Nau, Dana S, "Current trends in automated planning," *AI magazine,* Vol. 28, no. 4 (2007): 43-58., p. 43

[71] Ibid., p. 46

reliant on them and decrease their situational awareness.[72] A lack of situational awareness might not be a problem as long as a planning and scheduling system functions the way it should; however, problems may very well emerge when, at some point, the operator has to take control over or otherwise intervene in the system. Secondly, automated planning and scheduling may contribute to a de-skilling of individuals, both in terms of a loss in panning and organisational skill, and in terms of a loss in manual skills for the tasks that are being facilitated by automated planning techniques.

Thirdly, as automated planning and scheduling systems may influence decision-making,[73] the systems need to be trustworthy in order to take their suggested decisions seriously. If humans do not trust the system, it is less likely to be used. However, if the users trust the system too much they may never question its output and potentially miss errors made by the system.[74,75] The issue of trustworthiness relates to safety. For a system to be trusted, it is essential that it is believed to be safe. Especially high-level autonomous planning systems may be more susceptible to malicious actors due to a lack in human reasoning capabilities. This might impede the judgement to make a call on the trustworthiness of the received input, facilitating manipulation and hacking. Developers of autonomous planning systems should be aware of this risk in order to maintain users' trustworthiness in the system.

Furthermore, if the level of automation is high, it becomes unclear to whom the responsibility falls of the system's decisions made. This is discussed further in the part on "Machine learning" in subsection 5.1.2, and in the part on "Responsibility and accountability" in subsection 5.1.3.

Machine learning

This subsection details the ethical issues that arise from machine learning, which is an important technique in AI that has found widespread use in recent years. Machine learning is an efficient and effective way of programming based on statistics, as the programmer does not need to code every single action by hand (as is done with, for example, "if/then" statements in expert systems). Rather, the developer of the algorithm merely puts in a handful of guidelines and rules, after which the algorithm "learns" by itself.[76] A commonly used definition is provided by Tom M. Mitchell: "A computer program is said to learn from experience 'E', with respect to some class of tasks 'T' and performance measure 'P' if its performance at tasks in 'T' as measured by 'P' improves with experience 'E'."[77] A task may include for instance classification (classifying a value in a specific category) or regression problems (predicting a numerical value). The performance measure is examined based on test data: an algorithm is fed with input data, which is split into a set of training data and a set of test data. After the algorithm is trained, the test data is used to evaluate the accuracy of the algorithm. The experience focuses on the algorithm's learning process and can be divided into supervised and unsupervised learning.[78] Supervised learning is the most common form. An input is matched with an output and based on the

[72] Miller, Christopher A., Harry Funk, Robert Goldman, John Meisner, and Peggy Wu, "Implications of adaptive vs. adaptable UIs on decision making: Why "automated adaptiveness" is not always the right answer," In *Proceedings of the 1st international conference on augmented cognition*, pp. 22-27. 2005., p. 3

[73] Zimmerman, Terry, and Subbarao Kambhampati, "Learning-assisted automated planning: looking back, taking stock, going forward," *AI Magazine,* Vol. 24, no. 2, 2003, pp. 73-73.

[74] Miller et al. 2005, p. 2

[75] Dennis, Louise, Michael Fisher, Marija Slavkovik, and Matt Webster, "Formal Verification of Ethical Choices in Autonomous Systems," *Robotics and Autonomous Systems,* Vol. 77, 2016, pp. 1–14., p. 1

[76] Brynjolfsson, Erik, and Tom Mitchell, "What Can Machine Learning Do? Workforce Implications," *Science,* Vol. 358, No. 6370, 2017, pp. 1530–1534.

[77] Mitchell, Tom M., *Machine Learning*, McGraw Hill, New York, 1997.

[78] Brynjolfsson, Erik, and Tom Mitchell, "What Can Machine Learning Do? Workforce Implications," *Science,* Vol. 358, No. 6370, 2017, pp. 1530–1534.

test and training data the algorithm learns to match unseen inputs with unknown outputs. Therefore, it is generally used for classification and regression problems (e.g., detection of tumour, ranking house prices). Contrary to supervised learning, unsupervised learning has no previous knowledge of input-output relations. Instead, the algorithm is fed only input data in which the algorithm tries to find patterns. Therefore, this type is generally used for clustering (e.g., grouping customers). Sometimes a third distinction is made; reinforcement learning. Reinforcement learning is different from both supervised as unsupervised learning in the sense that it is constantly updated by a reward/punishment function. An initial state and a desired state are given, leaving the rest to the algorithm. The algorithm takes steps and based on whether it is rewarded or punished it will learn the best approach to reach the goal. Using this type of learning, breakthroughs in AI such as beating the Go master with AlphaGo have been accomplished.

Machine learning is becoming increasingly popular due to several factors, such as a boom in online data, low computational costs, and an improvement in learning algorithms.[79] In addition, ML algorithms are able to detect certain patterns in data humans are not able to, surpassing (certain) human capabilities. As the impact of ML algorithms in everyday life increases (e.g., decision on loans, job interviews), it is necessary to consider certain risks and worries that arise during the construction of these algorithms. Concerns that have arisen relate to ethical considerations such as fairness, interpretability (transparency, traceability, and explainability), reliability, responsibility and privacy. Although researchers have given these ethical issues increased consideration, it might be argued that their ethical analyses have not kept pace with the unabated development and widespread adoption of machine learning techniques.[80] The type of algorithm that causes most ethical issues are neural networks, due to specific characteristics that make them prone to bias and cause them to have an opaque character. The subsequent paragraphs in this subsection provide brief descriptions of each of the main ethical issues identified in relation to machine learning.

A common ethical issue in relation to ML is the potential for bias and discrimination as a result of unfair output by the algorithm. Unfair in this sense means that the algorithm favours a certain sex or race over another, which may negatively affect the possibilities of already disadvantaged and marginalised people.[81] There is a consensus that input data has a major influence on producing biased output.[82] The input data can be biased from the start,[83] correlations between features of the input data can be difficult to understand, or the algorithm exhibits a so-called uncertainty bias. This uncertainty bias arises when a minority in the data sample (less information and therefore less certainty) is disadvantaged because the algorithm prefers "to make decisions based on predictions about which it

[79] Jordan, M. I., and T. M. Mitchell, "Machine Learning: Trends, Perspectives, and Prospects," *Science,* Vol. 349, No. 6245, 2015, pp. 255–260.

[80] Thieltges, Andree, Florian Schmidt, and Simon Hegelich, "The devil's triangle: Ethical considerations on developing bot detection methods," *2016 AAAI Spring Symposium Series*, 2016.

[81] Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, *14*(3), 330-347.

[82] e.g. Kamishima, T., Akaho, S., & Sakuma, J. (2011, December). Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops* (pp. 643-650). IEEE.; Barocas, Solon, and Andrew D. Selbst, "Big Datas Disparate Impact," *Calif. L. Rev.,* Vol. 104, 2016, p. 671.; Brynjolfsson, Erik, and Tom Mitchell, "What Can Machine Learning Do? Workforce Implications," *Science,* Vol. 358, No. 6370, 2017, pp. 1530–1534.; Amini, Alexander, Ava Soleimany, Wilko Schwarting, Sangeeta Bhatia, and Daniela Rus, "Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure," *Proceedings of the 2019 AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES), 27-28 January, 2019 Honolulu, Hawaii, United States, AAAI/ACM,* 2019.)

[83] For more information see Barocas, Solon, and Andrew D. Selbst, "Big Datas Disparate Impact," *Calif. L. Rev.,* Vol. 104, 2016, p. 671.

is more confident."[84] Generalization may enlarge the uncertainty bias, as "minority records can be unfairly neglected."[85] Hence, reducing the features used by an algorithm, known as dimensionality reduction, may also increase inequality.

Besides the issue of problematic input data, the design of the algorithm may also raise fairness concerns. Burrell rejects the idea that algorithms are more objective than humans, because of the involvement of humans in the design in the algorithm.[86] She argues that "[t]his human work includes defining features, pre-classifying training data, and adjusting thresholds and parameters."

A second ethical issue with respect to ML is the general difficulty of explaining a ML-based system's output. Explainability is important for several reasons. First, users may be more likely to trust the system if they understand how the system reached its conclusion. Second, outcomes of the system are more easily justified when it is clear how it reached its conclusion. And third, one could argue that people have a right to explanations when the outcome of a system affects them. Not receiving an explanation may harm a person's agency and autonomy and could be considered a form of disrespect. An explanation allows someone to better challenge a decision made about them.

An outcome can be explained when the outcome is traceable, for which it needs to be transparent and interpretable. The question then is when a system can be considered interpretable. There is, however, no one exact definition of such.[87] The issue of opacity is generally related to neural networks,[88] but this issue is not necessarily *restricted* to this type of algorithm.[89]

Due to the opaque characteristics of neural networks, the problem of transparency, interpretability and explainability is strongest for these types of algorithms. When an algorithm is opaque it implies that it is unclear how a certain output is derived from an input.[90] This is partly due to the way algorithms tackle certain problems (e.g., image recognition, spam filtering), which is done differently than humans would. This difference makes it nearly impossible for humans to comprehend how algorithms come to their conclusion, independent of whether they have a high expertise on computer science. This problem raises the question whether we should focus on the *cause* of biased outcomes (i.e. why and

[84] Goodman, Bryce, and Seth Flaxman, "European Union regulations on algorithmic decision-making and a "right to explanation"," *AI Magazine,* Vol. 38, No. 3, 2017, pp. 50-57.

[85] (Kamishima, Akaho, & Sakuma, 2011, p. 2)

[86] Burrell, Jenna, "How the Machine Thinks: Understanding Opacity in Machine Learning Algorithms," *Big Data & Society,* Vol. 3, No. 1, 2015, p. 3.

[87] Lipton, Zachary C., "The Mythos of Model Interpretability," *Communications of the ACM,* Vol. 61, No. 10, 2018, pp. 36–43.; Doshi-Velez, Finale, and Been Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[88] E.g. Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199,* 2013.; Brynjolfsson, Erik, and Tom Mitchell, "What Can Machine Learning Do? Workforce Implications," *Science,* Vol. 358, No. 6370, 2017, pp. 1530–1534.; Litvinski, O. (2018). Algorithmic opacity: a narrative revue.

[89] E.g., Burrell, Jenna, "How the Machine Thinks: Understanding Opacity in Machine Learning Algorithms," *Big Data & Society,* Vol. 3, No. 1, 2015, p. 2053951715622512.; Lipton, Zachary C., "The Mythos of Model Interpretability," *Communications of the ACM*Vol. 61, No. 10, 2018, pp. 36–43.

[90] Burrell, Jenna, "How the Machine Thinks: Understanding Opacity in Machine Learning Algorithms," *Big Data & Society,* Vol. 3, No. 1, 2015, p. 2053951715622512.

how discriminative decisions arise), or rather on the *evaluation* of biased outcomes (i.e. decisions can be considered as discriminatory).[91,92]

Furthermore, the increase in data inhibits transparency of an algorithm in two ways. Firstly, it becomes more difficult to analyse an algorithm when there are more features to consider.[93] Secondly, due to an overload in features, dimensionality reduction is necessary in order to remain within the computational limits of an algorithm. This reduction, however, increases an algorithm's opacity as it might be unclear what features are ignored or combined with other features.

A third ethical issue regarding ML is reliability. Most ML algorithms are based on statistics. If an algorithm is 100% accurate on its test data, it is completely adjusted to the input data, including its outliers. This problem is known as overfitting, causing algorithms to be less accurate on unknown data. Therefore, training algorithms for 100% accuracy is practically unfeasible.[94] The algorithms therefore must make a trade-off between accuracy and robustness. Thieltges et al. include transparency in their trade-off and call it "the devil's triangle."[95] They argue that there is no general optimum between these ethical considerations. Complicating the algorithm makes it more complex and accurate, but less transparent. A transparent algorithm in turn may be easier to manipulate by exploiting exposed weaknesses in its design (i.e., "gaming the system"), thus causing a decrease in robustness.

A fourth ethical issue in relation to ML comprises its potential impacts on privacy and security. A common notion of privacy (i.e., "differential privacy") has generally been accepted,[96] advancing research relating to privacy concerns. Dwork and Roth explain differential privacy as a "paradox of learning nothing about an individual while learning useful information about a population."[97] Differential privacy is supposed to yield the same conclusion, independent of whether a certain individual was present in the data set. A problem for obtaining privacy in a data set is that algorithms have the ability to link features. Thus, in order to preserve anonymity, certain features may be removed from a data set. However, due to the ability to link features, the algorithm is still able to uncover unknown features, resulting in the so-called "red-lining effect."[98] Removing features is also in contrast with the efficiency of ML algorithms, as they improve with more available and workable data. This implies that algorithms work best when data sets are the least anonymous. This raises not only privacy

[91] For more information see: Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings. *Proceedings on privacy enhancing technologies*, *2015*(1), 92-112., who developed a tool to detect discriminatory results in shown advertisements, and Dwork, Cynthia, and Aaron Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science,* Vol. 9, No. 3–4, 2014, pp. 211-407., who look at a discriminative effect of particular decision, without focusing on the cause of this effect.

[92] Burrell, Jenna, "How the Machine Thinks: Understanding Opacity in Machine Learning Algorithms," *Big Data & Society,* Vol. 3, No. 1, 2015, p. 2053951715622512.

[93] Ibid.

[94] Brynjolfsson, Erik, and Tom Mitchell, "What Can Machine Learning Do? Workforce Implications," *Science,* Vol. 358, No. 6370, 2017, pp. 1530–1534.

[95] Thieltges et al. (2016, p. 253)

[96] e.g. Jordan, M. I., and T. M. Mitchell, "Machine Learning: Trends, Perspectives, and Prospects," *Science,* Vol. 349, No. 6245, 2015, pp. 255–260.; Papernot, Nicolas, Patrick Mcdaniel, Arunesh Sinha, and Michael P. Wellman, "SoK: Security and Privacy in Machine Learning," *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, 2018.

[97] Dwork, Cynthia, and Aaron Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science,* Vol. 9, No. 3–4, 2014, pp. 211-407.

[98] (Kamishima, Akaho, & Sakuma, 2011, p. 644; see also Dwork, Cynthia, and Aaron Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science,* Vol. 9, No. 3–4, 2014, pp. 211-407.(p.7)

concerns, but also questions in the ownership of data[99] and security problems.[100] These trade-offs between privacy and accuracy and between privacy and security are difficult to assess.

A fifth and final ethical issue with regard to ML is its potential effects on responsibility and accountability. Mitchell's definition specifically focuses on the algorithm and excludes surrounding social components such as developers and users of the algorithm. This exclusion may give the impression that human components are a priori excluded from the process, and therefore not or less responsible for the algorithm's outcomes and consequences. This may result in a neglect of social and ethical considerations from the machine learning domain. Cerna Collectif (2018) explains the difficulty of assigning responsibility.[101] They argue that, generally, the designer of the system should be responsible when the system is flawed, and the user should be responsible when he or she abuses the system. Machine learning needs training however, and this trainer could also be at fault (e.g., bad training data). Not infrequently do machine learning systems update themselves by data received from the users (e.g., recommendation algorithms on social media platforms). The trainer and user become one in this case, complicating responsibility issues. In addition, Matthias (2004) has raised the contemporary concern of "responsibility gap."[102] Such a gap exists when the manufacturer of a machine cannot be held accountable to the machine's reaction, due to a loss of "control over the device."[103] Matthias argues that developers may be regarded as the "*creator* of software organism", that once "released" develop plans and actions outside the control of the programmer. Therefore, he argues that programmers cannot be held morally responsible for the machine's actions.

Machine ethics

Machine ethics (also known as *artificial morality*, *machine morality*, and *computational ethics*) is an emerging field of study at the intersection of AI and ethics aimed at investigating ways to implement ethical decision-making faculties in machines (e.g., computers, robots).[104] There are two main reasons for pursuing this area of inquiry. First, it is hoped that computational modelling of human morality will help to achieve a better understanding of human morality. Second, as machines are becoming ever more autonomous and increasingly taking on human tasks and operating among humans, equipping them with capabilities to compute ethical decisions is, at least in contexts where moral dilemmas are likely to occur, seen as an indispensable requirement. Currently, there exist a fair number of approaches to the creation of what we may call "ethical reasoning systems". These approaches can be divided into "top-down" approaches, which involve the explicit programming of an ethical theory into a machine; "bottom-up" approaches, which progressively build up an ethical framework through the use of case-based reasoning or learning-based methods; hybrid approaches, which combine the previous two approaches; psychological approaches; which seek to mimic the cognitive processes of human ethical decision-making; and artificial general intelligence (AGI) ethics proposals, which are

[99] Jordan, M. I., and T. M. Mitchell, "Machine Learning: Trends, Perspectives, and Prospects," *Science,* Vol. 349, No. 6245, 2015, pp. 255–260.
[100] Mitchell, T. M. (2006). *The discipline of machine learning* (Vol. 9). Pittsburgh, PA: Carnegie Mellon University, School of Computer Science, Machine Learning Department.
[101] Collectif, Cerna. "Research Ethics in Machine Learning," PhD diss., CERNA; ALLISTENE, 2018.
[102] Matthias, Andreas, "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata," *Ethics and Information Technology,* Vol. 6, No. 3, 2004, pp. 175–183.
[103] Matthias, Andreas, "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata," *Ethics and Information Technology,* Vol. 6, No. 3, 2004, pp. 175–183(176).
[104] Allen, Colin, Wendell Wallach, and Iva Smit, "Why Machine Ethics?," *IEEE Intelligent Systems*, Vol. 21, No. 4, 2006, pp. 12–17.

aimed at constraining the behaviour of advanced artificial general intelligence systems.[105] The next few paragraphs discuss five sets of ethical issues in relation to ethical reasoning systems: (1) issues stemming from the very nature of ethics, (2) issues arising from the potential for system failure and corruptibility, (3) issues in relation to the risk of creating moral patients, (4) issues resulting from unequal distribution of benefits, and (5) issues with regard to moral responsibility and accountability.

The first set of ethical issues arise from the fact that there are many unsolved problems in ethics and that there may exist genuine moral dilemmas. Since the human intuitions that ground ethical theories are unsystematic at their core,[106] it may not prove feasible to develop ethical reasoning systems that use algorithms to consistently arrive at a single best moral judgment that accords with these intuitions. Especially if one is after a "top-down" approach to machine ethics, it is very difficult to specify an ethical theory or framework that is consistent with human intuitions and acceptable to everyone. This is evidenced by an academic literature that lacks consensus as to which ethical theory best represents human morality, and is rife with critical discussion of consequentialist[107], deontological[108] and other theories—including those that try to amenably synthesise the insights of the first two. The deployment of an ethical reasoning system using an overarching ethical theory that is not entirely consistent with human intuitions in a wide range of situations and does not receive broad public support is obviously going to be highly problematic from an ethical standpoint. To be sure, there are also "bottom-up" approaches to developing moral reasoning systems based on, for example, machine learning and case-based reasoning, but these too have problems, as explained further on.

Besides the challenge of creating a comprehensive overarching ethical theory for use in AI systems, the prospect of ethical reasoning by computational systems is further challenged by the notion of *value pluralism*.[109] This issue is very much related to, but can also be seen as distinct from, the aforementioned issue. Value pluralists believe that it is impossible to reduce all moral values to a single value, such as pleasure, welfare or happiness.[110] They further hold that, as a result of the irreducibility of *incommensurable* values, unresolvable moral dilemmas are likely to emerge in cases where multiple

[105] Brundage, Miles, "Limitations and risks of machine ethics," *Journal of Experimental & Theoretical Artificial Intelligence*, Vol. 26, No. 3, 2014, pp. 355–372.

[106] In moral psychology, a consensus seems to be emerging that, depending on the specifics of a particular situation and factors such as cognitive load, humans either make use of an intuitive moral cognitive system or a more deliberate moral cognitive system. It has been argued that this so-called "dual-process model" approximately maps onto the distinction in moral philosophy between deontological (means-based) and consequentialist (ends-based) theories. The dual-process nature of human cognitive processing may explain the persistence of moral problems and the difficulty of arriving at a well-specified moral theory without exceptions. Cushman, Fiery, Liane Young, and Joshua Greene, "Multi-system moral psychology," In *The moral psychology handbook*, J. M. Doris & the Moral Psychology Research Group, Eds., New York, NY: Oxford University Press, 2012, pp. 47–71.

[107] In the ethical literature, consequentialist theories have been variously criticised for not being able to sufficiently account for the moral value of one's social commitments to, for example, friends and family, as well as one's life projects; for putting excessive demands on persons to contribute to the welfare of others; for arriving at unacceptable conclusions in some cases; and for failing to sufficiently recognise individual rights, distributive justice considerations, and the separateness of persons. Brundage, Miles, 2014, op. cit.

[108] In the ethical literature, deontological theories have been variously criticised for potentially producing catastrophic results in cases where there are extreme trade-offs to be made between the interests of few and the interests of many; for their general inability to adequately deal with conflicts between duties; and for their collapsing into consequentialism given that an actor who opposes generating harm A is rationally committed to reducing the amount of B in their environment. Brundage, Miles, 2014, op. cit.

[109] Brundage, 2014, op. cit.

[110] Value pluralism can be contrasted with *valuemonism*, the belief that all values can be ordered and reduced to a single value, such as the *human good*, meaning that all value conflicts are ultimately resolvable.

values compete with one another.[111] If value pluralism is true, then it is unreasonable to expect an ethical reasoning system to ever be able to resolve every complex moral dilemma it encounters, since oftentimes no single best solution would exist.[112,113] Incorporating the notion of value pluralism in ethical reasoning systems may entail unpredictable behaviour by the AI system or paralysis in cases where value trade-offs are expected,[114] thus potentially jeopardising peoples' safety. In addition, any heuristic used to overcome truly unresolvable moral dilemmas in a particular fashion may, if employed on a large scale, become highly influential. It has been argued that this could result in a kind of "value imperialism", which could negatively affect cultures (especially those that were not involved in the development of the system) and degrade cultural autonomy.[115,116] This latter issue may also apply to the selection of an overarching ethical theory for use in an ethical reasoning system.

Further challenges resulting from the very nature of ethics that may inhibit the creation of ethical reasoning systems include longstanding unresolved problems in ethics, such as: population ethics; issues related to the possibility of infinite value; small probabilities of enormous amounts of value; the relationship between theoretical virtues and intuitions; and moral uncertainty.[117] Another potential problem worth mentioning is that formulating ethical values as quantifiable parameters computable by an ethical reasoning system may prove to be a very difficult and contentious task.

A second set of ethical issues with regard to ethical reasoning systems comprise the risks to safety, security and other ethical values as a result of these systems' potential for failure and corruptibility. The potential for system failure arises from the computational and knowledge limitations that are present when bounded agents operate in complex environments. First, there is a significant risk of incorrect input into the ethical reasoning system. For their proper functioning, an ethical reasoning system depends on being supplied with relevant and accurate information about the environment in which it operates.[118] Even the most advanced systems may reach false conclusions on matters of great ethical significance if the inputs are wrong. In some cases, it may be near impossible to supply an ethical reasoning system with each and every piece of information that may have bearing on its ethical decision-making. Second, there is the intractability of computing, on the basis of the inputs, the ethically relevant implications if a large number of agents or actions, or a long time-horizon is involved. In difficult (often socially complex) cases, this may lead even the best ethical reasoning system that is given perfect information to reach ethically unacceptable conclusions.[119]

These issues will be at play—albeit perhaps to varying degrees—regardless of which specific ethical theory lies at the core of the ethical reasoning system. Systems that use a "bottom-up" approach (e.g.,

[111] Stocker, Michael, "Abstract and concrete value: Plurality, conflict and maximization," in *Incommensurability, Incomparability and Practical Reason*, R. Chang, Ed., Cambridge, MA, USA: Harvard Univ. Press, 1997.

[112] Brundage, 2014, op. cit.

[113] Anderson, Susan Leigh, "Machine metaethics," in *Machine Ethics*, M. Anderson and S. Anderson, Eds., New York, NY, USA: Cambridge University Press, 2011, pp. 21–27.

[114] Brundage, 2014, op. cit.

[115] Cave, Stephen, Rune Nyrup, Karina Vold, and Adrian Weller, "Motivations and Risks of Machine Ethics," *Proceedings of the IEEE*, Vol. 107, No. 3, 2019, pp. 562–574.

[116] Value imperialisation has been defined as "the universalization of a set of values in a way that reflects the value system of one group". Cave, et al., 2019, op. cit.

[117] Please note this listing is not exhaustive. Find more unsolved issues here: Crouch, Will, "The most important unsolved problems in ethics," 2012. http://blog.practicalethics.ox.ac.uk/2012/10/the-most-important-unsolvedproblems-in-ethics-or-how-to-be-a-high-impact-philosopher-part-iii/

[118] Cave et al., 2019, op. cit.

[119] Allen, Colin, Gary Varner, and Jason Zinser, "Prolegomena to Any Future Artificial Moral Agent," *Journal of Experimental & Theoretical Artificial Intelligence*, Vol. 12, No. 3, 2000, pp. 251–261.

using machine learning), however, may also be vulnerable in the sense that they may infer morally unacceptable ethical principles from improper inputs and supervision.[120,121] It is impossible to supply such a system with an infinite number of perfect training examples to guarantee its flawless operation.

There is also the issue of what standard of fallibility society should want ethical reasoning systems to adhere to. For these systems, simply adhering to human standards may not be good enough. If an individual system makes few and minor mistakes at an acceptable level for humans, it might still mean that the very same mistakes by a large number of such a system may amount to an unacceptable problem in the aggregate.[122] Additionally, when ethical reasoning systems fail, their failures may have a very high impact since machines often fail in unpredictable and difficult-to-manage ways.[123]

In a world where all humans are well-intentioned, our discussion of ethical issues in relation to the potential for system failure would end here. Unfortunately, however, we have to account for the potential that ethical reasoning systems can purposely be turned into unethical systems. Various authors have noted how such systems are easily corruptible by hackers or malicious designers or trainers (as well as through simple coding errors).[124,125] Such risks may be compounded if malicious ethical reasoning systems also possess a powerful capacity to generate deceptive or manipulative explanations for their actions.[126]

A third group of ethical issues relate to the risk of creating *moral patients*. The term moral patient commonly refers to any entity (e.g., humans, animals, species, ecosystems) whose interests are thought to matter (e.g., because they can experience pain or suffering) and who should not be harmed or wronged absent reasonable justification. While it may seem unlikely that within the next 20 years machines will have developed the complex cognitive capacities that allow for phenomenological consciousness and the experiencing of pleasure, pain, and suffering, their moral patiency may yet arrive through a different route. The advanced ethical reasoning systems of some future machines may simply possess self-reflexive qualities that make these machines *appear* as though they are agents that have intentionality.[127] This may lead humans to grant them status as *moral agents*, a term used to identify those entities that bear moral responsibilities towards moral patients. By virtue of their moral agency (since all moral agents are also moral patients), these systems would then also have a status as moral patients.

The emergence of machines that can be considered moral patients could create new moral obligations for humans to take seriously the interests of such machines. It has been argued that these duties could potentially have enormous costs and constrain humans in significant ways. For instance, humans might

[120] Brundage, Miles, 2014, op. cit.
[121] Allen, Colin, and Wendell Wallach, *Moral Machines: Teaching Robots Right from Wrong*, London, U.K.: Oxford University Press, 2009.
[122] Cave et al., 2019, op. cit.
[123] Athalye, Anish, Logan Engstrom, Andrew Ilyas, and Kevin Kwok, “Synthesizing Robust Adversarial Examples,” Cornell University arXiv.org, 2018. https://arxiv.org/abs/1707.07397
[124] Vanderelst, Dieter, and Alan Winfield, “The Dark Side of Ethical Robots,” Cornell University arXiv.org, 2016. https://arxiv.org/abs/1606.02583
[125] Charisi, Vicky, Louise Dennis, Michael Fisher, Robert Lieck, Andreas Matthias, Marija Slavkovik, Janina Loh (Sombetzki), Alan F.T. Winfield, and Roman Yampolskiy, “Towards moral autonomous systems.” Cornell University arXiv.org, 2017. https://arxiv.org/abs/1703.04741
[126] Cave et al., 2019, op. cit.
[127] As noted by Cave et al., for some, moral status is already grounded by a capacity for self-awareness, the ability to reflexively represent oneself. Currently, a kind of self-monitoring and self-representation already exists in some algorithms that use neural networks with hierarchical layers where the higher levels predict the chance of success for the lower layers. Cave et al., 2019, op. cit.

have to respect the right of these machines to exist and to not be turned off, as well as their right to autonomy, and they might not be able to employ the machines as mere tools or slaves.[128] Humans might also have to share with the machines some of their privileges, such as the right to vote in elections.[129] Perhaps fairness would even dictate that the machines be given their own homeland. All of this may put pressure on the multitude of responsibilities humans currently have to one another and their environment (e.g., human rights, animal rights, distributive justice), which, as it stands, they already have trouble fully meeting.

A fourth set of ethical issues concern distributive justice. AI systems equipped with ethical reasoning capabilities *may* prove to be of great value for individuals and groups, and they *may* lead to substantial improvements in overall human wellbeing. Yet, their deployment may also be financially costly, and therefore it might not be feasible to introduce these systems in each and every context where their use could prove beneficial. This could mean that not everyone may stand to benefit equally from the technology, which in turn could, in some cases, raise distributive justice concerns. One could ask, perhaps, whether in particular contexts all humans may have a right to receive assistance from advanced artificial agents equipped with well-functioning ethical reasoning systems, if such systems have proven to be practically feasible. This may also be an issue of power relations if it turns out that only large and powerful entities such as militaries and large corporations possess sufficient resources and motivation to develop ethical reasoning systems and that they are disinclined to relinquish control and diffuse the benefits of these technologies beyond their own spheres of operation.

A fifth and final group of ethical issues in relation to machine ethics comprise the argument that ethical reasoning systems may undermine human moral responsibility. *(More on the ethical issues in relation to responsibility and accountability in subsection 5.1.3.)* Cave et al. (2019) claim that such systems may do so in three ways: they may weaken (1) humans' capacity to make proper moral judgements, (2) their ability and willingness to use this capacity, and (3) their ability and willingness to assume responsibility for ethical decisions and outcomes.[130] All three aspects are said to result from a so-called "automation paradox", which refers to the problem that labour-saving automated machines tend to (1) compensate for human incompetence, (2) erode existing human skills, and (3) fail in the most pressing and unusual situations that are likely to catch humans off-guard and ill-prepared.[131]

With respect to the first strand of the automation paradox, Cave et al. (2019) argue that humans may not be able to properly develop their moral reasoning skills as a result of using automated systems that either take over ethical decision-making from humans entirely or assist humans in making ethical decisions.[132] Further, in relation to the second aspect, they point out that moral reasoning is also a skill that humans need to continually practice so as to prevent it from slowly deteriorating.[133] The erosion of moral reasoning skills due to using moral reasoning systems is held be most severe in cases where the entire ethical decision-making process is automated. Finally, with regard to the third aspect, the

[128] Cave et al., 2019, op. cit.

[129] Cave et al., 2019, op. cit.

[130] Cave et al., 2019, op. cit.

[131] Harford, Tim, "Crash: How Computers are Setting us up for Disaster," *The Guardian*, 2016. https://www.theguardian.com/technology/2016/oct/11/crash-howcomputers-are-setting-us-up-disaster

[132] As an example, Cave et al. argue that the use of a hypothetical medical robot that can make ethical decisions might result in a situation where medical staff has not been able to develop adequate judgment and sensitivity to decide when patient autonomy should be sacrificed against patient well-being in cases where a patient does not comply with medical advice.

[133] Cave et al. present its emphasis in the education and socialisation of children, as well as in professional education (e.g., in medicine), as evidence that moral reasoning is indeed a skill that is not innate but has to be developed and maintained.

authors state that while machines with ethical reasoning systems may deal with day-to-day ethical issues in a successful manner, there may occasionally be very difficult cases that the machine identifies as being beyond its capabilities. Decision-making for such cases will then be referred back to humans. However, exactly these cases may prove to be extremely challenging for humans, as they are likely to be novel and complex, and decisions on them may need to be made quickly (such as in autonomous driving). Considering also the first and the second aspect of the automation paradox, humans may be very ill-prepared to deal with such difficult cases, which may have considerable consequences for human safety, justice and well-being.

Beyond these three issues, there are a few other issues that relate to responsibility. One is that at least for important ethical decisions, humans would want machines to be able to explain, in human-understandable terms, how they reached those decisions. This, however, poses significant technical challenges, especially in relation to neural network-based systems. *(More on the ethical issues in relation to transparency and explainability in subsection 5.1.3.)*

Such technical challenges aside, it may be possible that in certain cases the reasoning processes of ethical machines are complex to such a degree (e.g., in cases that involve a large number of value trade-offs) that no human will ever be able to understand them. In these cases, humans will be unable to evaluate whether ethical reasoning systems made the right calls for the right reasons, thus losing their ability to hold the machines, or any humans associated with their development or use, to account. Additionally, as the complexity of such systems moves beyond human capabilities for understanding them, there may be difficulties regarding the allotment of trust to these systems, which may have implications in terms of well-being and safety.

Related to this, Cave et al. (2019) argue that in a world in which each and every ethical decision is being made by machines skilled at solving even the most challenging ethical problems, humans may stop using their moral faculties altogether, and consequently would not even know what it meant for the ethical reasoning systems to fail. They contend that “passing enough consequential ethical decisions over to machines too complex for us to understand could therefore pose a risk to the entire system of moral reasoning, reason-giving and responsibility.”[134]

5.1.3. Ethical issues with regard to general implications and risks

In this subsection, we describe the main ethical issues with regard to the general implications and risks of AI technology. For each ethical principle and type of harm that we have identified as being implicated any potential negative consequences of the development and use of AI technology, detail the ways in which harm can potentially occur. We focus on *autonomy and liberty*, *privacy*, *justice and fairness*, *responsibility and accountability*, *safety and security*, *dual use and misuse*, *mass unemployment*, *transparency and explainability*, and *other potential harms*, respectively.

Autonomy and liberty

Autonomy and liberty (or freedom) are often identified as values that could be threatened by the indiscriminate use of AI. In this section, we will first discuss the nature of these two values, and their relation to each other. We will then discuss the different ways in which they could be harmed by AI, as well as ways in which AI can also support them.

We will now analyse the concept of autonomy, followed by the concept of liberty, and we will also explore the relation between them. The term “autonomy” comes from the Greek word αὐτόνομος,

[134] Cave et al., 2019, op. cit., p. 572.

which means self-rule (from the Greek word autos meaning "self" and nomos meaning rule or law. It was originally used to refer to self-ruling city states, but is now primarily used to refer to persons. An autonomous person is a person who is self-governing or capable of self-rule.

To be self-governing, two conditions must be fulfilled. First, one must be able to make decisions based on values, principles, desires, and deliberations that are one's own, and that are not the result of manipulation or coercion by others.[135] This requires that one's values, desires, etc. are authentic: they are ones that are formed by one's own volition and deliberations, without undue influences by others, and endorsed by one upon further reflection and evaluation. A second condition that must be fulfilled is that one has the capacity to act competently on one's authentic values and desires. This requires that one has capacities for rational thought and for self-control, and is free of pathologies like systematic self-deception.[136]

Autonomy has become an important ideal for persons in the modern era. It is, to be sure, an ideal that may never be reached fully, since it appears that our values and desires are always influenced and manipulated by others to some extent, and we are not fully rational beings who are capable of full self-control. However, most adult human beings are capable of basic autonomy, which is the state of being "responsible, independent and able to speak for oneself".[137] Autonomy is considered to be a precondition for moral and legal responsibility, and for political equality. Persons who are not considered autonomous include children and people with severe mental disabilities that impair their capability of autonomous judgment. They are not considered to be responsible for their actions, and they enjoy more limited rights than autonomous persons and can become subjected to paternalism.

Autonomy is usually distinguished from freedom or liberty, in that freedom is usually defined in relation to one's ability to act without constraints, whereas autonomy concerns one's ability to make independent decisions based on authentic values and desires. In principles, one can be had without the other, as can be illustrated by the following two cases. The first is that of Nelson Mandela, the South-African anti-apartheid revolutionary, who spent 27 years in prison. During his time in prison, Mandela clearly enjoyed very little freedom, in that he was prevented from performing many actions that he might have wished to perform. Yet, during this period, he maintained full autonomy, in that he maintained an unbroken spirit, retaining the belief in the values and principles that he stood for, and the desire to end apartheid. The second is that of Caligula, the ancient Greek emperor and tyrant, who enjoyed vast powers and could do anything he wanted with impunity. Caligula thus enjoyed unprecedented liberties. However, he was by all accounts not an autonomous person, suffering from narcissism and paranoia, and possibly other mental diseases which were not diagnosed at the time.

The relationship between autonomy and liberty becomes muddier when one considers that liberty, on most philosophical conceptions, does not only concern the ability to act without *external* constraints, but also the ability to act without *internal* constraints. A person who is in prison is unfree, but a person with agoraphobia is similarly unfree. Both are constrained in going outside, the first by external constraints, and the second through internal constraints. At the same time, one could claim that a person with agoraphobia is not fully autonomous, since some of his or her desires and decisions are not fully authentic and rational.

[135] Dworkin, Gerald, *The theory and practice of autonomy*, Cambridge University Press, New York, 1988., 61f; Arneson, Richard, "Autonomy and Preference Formation," 1991, in Coleman, Jules L. and Allen Buchanan, eds, *In Harm's Way: Essays in Honor of Joel Feinberg*, Cambridge University Press, Cambridge, 1994, pp. 42–73.
[136] Christman, John, "Autonomy in Moral and Political Philosophy," *Stanford Encyclopedia of Philosophy*, Edward N. Zalta, January 9, 2015. https://plato.stanford.edu/entries/autonomy-moral/.
[137] Ibid.

To further complicate matters, Isaiah Berlin has famously distinguished two senses of freedom or liberty, which he named negative and positive liberty.[138] Negative liberty is freedom from external constraints. Positive liberty is the ability to self-control or self-mastery. This is the ability to overcome internal constraints and to make autonomous choices. Berlin himself likens positive liberty to autonomy, and many commentators since have equated the two. So, following Berlin, autonomy is actually a *component* of liberty, rather than a complementary value.

In practice, however, there are good reasons to keep distinguishing autonomy from liberty, because from an ethical point of view the two concepts impose different moral requirements on actors. Freedom rights, as defined in for example the Universal Declaration of Human Rights or the European Charter, typically refer to rights to perform certain actions without *external* constraints. For example, they are rights to freedom of expression, freedom of assembly, or and rights not to be subjected to arbitrary arrest, detention or exile. Rights to freedom from *internal* constraints are normally not identified as such, because third parties normally do not have the power to cause these internal constraints to be in place in person. Even when they do, as for example when a drug dealer contributes to someone's addiction, this is not normally seen as a violation of freedom rights, but rather as a harm to health. So, from a practical point of view, moral principles relating to freedom tend to apply to negative freedom and involve external constraints on freedom.

Respecting autonomy in others requires different moral actions. It requires that one does not manipulate their desires, control their thoughts, undermine their capacity for self-control and independent deliberation, or coerce their decisions. These are largely actions that are different from those involved in the imposition of external constraints to freedom. They may, however, accompany each other, as in enslavement, which may involve both psychological manipulation and brainwashing (undermining autonomy) and confinement, enchainment and physical punishment (limiting freedom). Bridging freedom and autonomy, the principle of freedom of thought and religion, found in both the Universal Declaration of Human Rights and the European Charter, is enumerated as a freedom right, but also relates to autonomy. Freedom of thought may be limited in two fundamental ways. First, it may be limited by curtailing the liberties of individuals, for example by limiting freedom of expression, of assembly, and of religious service. These actions do not directly affect autonomy, but they limit freedoms and may affect autonomy indirectly. Second, it may be limited through compulsory re-education programs and camps in which people are actively taught to have different ideas, values and preferences. In such cases, autonomy is affected directly.

Let us now discuss the ethical impacts of AI on autonomy and liberty, starting with the impact on autonomy. AI can negatively affect autonomy in at least three ways. First, AI is a technology capable of making decisions and acting on them. This undermines the autonomy of persons if they would otherwise have made the decision, and are prevented from doing so without their consent. Second, AI can recommend decisions to persons in a context that they have not fully consented to but that leaves them little choice but to follow the recommendation. Third, AI can be designed to explicitly or subliminally influence and condition people's desires, values and beliefs.

Decisions made by AI, first of all, can undermine autonomy by keeping people from thinking for themselves and making their own decisions. Even if the actions prescribed by these decisions are carried out by people themselves, their autonomy is diminished if they do not get to decide. Consider a hypothetical example, in which people carry with them at all times a highly intelligent AI system that constantly collects and processes information about them and their environment, and decide for them

[138] Berlin, Isaiah, Two Concepts of Liberty, 1969, in Berlin, Isaiah, *Four Essays on Liberty*, Oxford University Press, Oxford, p. 118-172.

what to eat, what to do, where to go, and even what their overall life goals should be. Such a system clearly undermines autonomy in a very substantial way. Obviously, though, the more restricted the scope of an AI system is, and the most innocuous the decisions it makes, the less autonomy will be undermined.

Even if the role of an AI system is only to recommend certain decisions, rather than make them, autonomy can still be curtailed as a result. People may come to trust decision support systems, because they believe that they are capable of making choices that are better, or at least as good, as they would make them, and they may like the delegation of responsibility to such systems, relieving them of the burden of choice. In other cases, they may not trust or like the system, but are expected within their profession, to consult the system, as sometimes applies to medical or legal expert systems. If the system is held by others to give reliable advice, it may be difficult for professionals to ignore the system and depend on their own judgment. Thus, their decisional autonomy is limited as well.

The third way that was identified in which AI systems can undermine autonomy is by explicitly or subliminally influencing values and desires. There are at least three ways in which this may happen. A first is through targeted messaging and advertising based on advanced personal profiles. Advanced personal profiles are digital profiles of persons that contain not only demographic and socioeconomic data but also data about their (online) behaviours, such as websites they have visited and products they have purchased. Such data may be used to make inferences about people's needs, interests and desires, and this may be used to offer them content (advertising, news and information) that are likely to be of interest to them.

A major way in which such content is offered to people is through personalized recommender systems, which are systems that select and recommend content to individual in which they are expected to have an interest. This type of targeted messaging potentially undermine autonomy by presuming to know what people's preferences are, without asking them explicitly, and then tailoring their information environment as a result. In doing so, it may moreover reinforce certain preferences and beliefs at the expense of others. Even more so, they may limit our autonomy by only reinforcing those preferences, desires and beliefs that we have had in the past, and limiting our exposure to substantially new content [139] and by manufacturing and engineering new needs and desires. These processes are especially worrisome when they affect one's exposure to a diversity of information and opinions, as it may place us in "filter bubbles" that only reinforce one's present beliefs and opinions).[140] These systems are moreover not neutral in that there is usually a commercial interest behind recommender algorithms, which tends to favour recommendations that can lead to profit.

Targeted messaging is not only used for recommending content or for creating online filter bubbles. It is also used for nudging. A nudge is a stimulus that influences people's choices and behaviours in predictable ways without forbidding choices or changing people's economic incentives.[141] Nudges can be very simple stimuli, such as lines on the pavement that suggest where people should walk, or the display of products at eye level on a counter. However, with the advent of AI, nudging has taken the form of textual and non-textual messages sent to you by health and lifestyle apps and everyday products that are part the Internet-of-Things, such as your fridge suggesting that it needs restocking

[139] Nitzberg, Mark, Olaf Groth, and Mark Esposito, "AI Isn't Just Compromising Our Privacy-It Can Limit Our Choices, Too," *Quartz*, December 13, 2017. https://qz.com/1153647/ai-isnt-just-taking-away-our-privacy-its-destroying-our-free-will-too/

[140] Pariser, Eli, *The Filter Bubble: What the Internet Is Hiding from You*, Penguin Books, London, 2011.

[141] Thaler, Richard H., and Cass R. Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness*, Penguin Books, New York, 2008.

and the thermostat glowing so as to indicate that your energy use is above normal. While such nudges may have benefits, social scientist Joseph Coughlin (2017) has warned for a 24/7 nudge economy in which we are bombarded with messages to influence our decisions and behaviours.[142] Nudging at this scale may well limit our autonomy, by constantly influencing and steering our choices and decisions.

A second, more sophisticated way of in which AI systems can be used to influence our values and desires is through their use in psychographic modelling.[143] Psychographic modelling goes beyond demographic, socioeconomic and online behavioural data to include psychographic data, which relates to values, beliefs, emotions, personalities, interests and lifestyles. Social media, in particular, contains a lot of data from which psychographic information can be derived. Psychologists work with IT specialists to segment populations in this way and build complex psychographic profiles of different groups. Such profiles can reveal vulnerabilities of these groups and be used for messaging that influences their values and preferences. There is evidence that this type of targeted messaging can be highly effective in influencing people's preferences and opinions.[144] It is mostly used in advertising, but has recently also been used by Cambridge Analytica for political messaging, notably in the 2016 U.S. elections.[145]

A third and final way in which AI can influence our values and preferences is by inducing dopamine-driven feedback loops. Research in neuroscience has shown that rewarding social stimuli activate particular dopamine pathways in the brain that generate pleasurable feelings.[146] When these pathways are activated frequently by the same behaviour or stimulus, resulting in rewarding feelings, the association between the stimulus or behaviour and reward is strengthened. This can induce addictive behaviour in people, in which they cannot stop from performing certain behaviours in order to get a dopamine boost. According to former employees of firms like Facebook and Google, this knowledge has been exploited by tech companies to get people addicted to social media, apps and games through likes, push messages, and other manufactured compulsion loops.[147] Specialized companies exist, such as Dopamine Labs, to exploit neuroscientific insights in order to "hook" users to digital media. Not all of these efforts are AI-driven, but AI is being used to bring them to the next level. Clearly, these practices undermine autonomy by limiting the authenticity of desires and the rational foundation behind people's choices.

Next to all these negative impacts, AI can also impact autonomy positively. As we have argued before, AI can be a double-edged sword with respect to autonomy. By taking decisions away from us, it can diminish our autonomy by depriving us from the opportunity to make these decisions ourselves, but it

[142] Coughlin, Joseph F., "The 'Internet of Things' Will Take Nudge Theory Too Far," *Big Think*, March 27, 2017. https://bigthink.com/disruptive-demographics/the-internet-of-things-big-data-when-a-nudge-becomes-a-noodge

[143] Liu, Hui, Yinghui Huang, Zichao Wang, Kai Liu, Xiangen Hu, and Weijun Wang, "Personality or Value: A Comparative Study of Psychographic Segmentation Based on an Online Review Enhanced Recommender System," *Applied Sciences*, Vol. 9, No. 10, 2019, pp. 1992, DOI:10.3390/app9101992.

[144] Matz, Sandra C., Michal Kosinski, Gideon Nave, and David J. Stillwell, "Psychological targeting as an effective approach to digital mass persuasion," *Proceedings of the National Academy of Sciences of the United States of America* Vol. 114, No. 48, 2017, pp. 12714–12719.

[145] Wade, Michael, "Psychographics: the Behavioural Analysis That Helped Cambridge Analytica Know Voters' Minds," *The Conversation*, March 21, 2018. http://theconversation.com/psychographics-the-behavioural-analysis-that-helped-cambridge-analytica-know-voters-minds-93675

[146] Krach, Sören, Frieder M. Paulus, Maren Bodden, and Tilo Kircher, "The Rewarding Nature of Social Interactions," *Frontiers in behavioral neuroscience,* Vol. 4, p. 22, 2010.

[147] Tiffany, Kaitlyn, "A Timeline of High-Profile Tech Apologies.," *Vox*, July 26, 2019. https://www.vox.com/the-goods/2019/7/26/8930765/tech-apologies-former-facebook-google-twitter-employees-list

can also disburden us by taking away unimportant decisions, thus enhancing our autonomy by giving us more time and energy to focus on important decisions. It can also be used to train our mind, enhancing our potential for deliberation and self-understanding, and correcting our cognitive biases, and to nudge us to make healthy choices and reign in our impulses and bad habits, thus enhancing our autonomy in the long run.

To protect human autonomy, various requirements have been proposed as mandatory for AI systems, including human oversight, human-in-command, human-in-the-loop and meaningful human control.[148] These are quite different notions, but they have in common the idea that humans should always be in control of AI, either by being able to assess the operation and consequences of an AI system and the value and necessity of its use, or by controlling its decision-making process by being involved in it or having the ability to intervene. It has also been proposed that the selective exposure to content brought about through personalisation and recommender systems should be counteracted through algorithms and methods that promote more diverse exposure to information and break so-called filter bubbles.[149] Psychographic modelling and feedback loops have been severely criticized, but they are still in use in the industry, and it should perhaps be considered if these approaches should be outlawed.

Having discussed the main ethical issues in terms of autonomy, let us now turn to the impact of AI on liberty. AI can limit human freedom in two basic ways. First, AI systems can take automated actions that impose constraints on humans, limiting their abilities to act. Second, AI systems can provide information to third parties that can help them impose constraints on freedoms of individuals. We will discuss these two types of limitation in order.

Actions that AI systems can take are either informational or physical. Informational actions are actions defined over digital information, and physical actions are actions that result from AI systems being equipped with actuators or being coupled with machines whose operations they can control. Both types of actions can be used to limit human freedoms in a direct way. Informational actions can be used to limit freedom of expression, limit online actions by human users, and limit access to online resources. AI is already being used on a large scale to limit freedom of expression by identifying and removing content that violates certain standards. While such censorship can be justified (e.g., removal of terrorism-related content), it can also be used for unjustified censorship by authoritarian regimes. AI can also be used to limit online freedoms by restricting access to certain sites, not allowing certain transactions to take place by certain individuals, or restricting access to online resources by users who fit a certain profile.

AI systems can also physically limit the freedoms of individuals. AI-controlled security systems may automatically close gates and physically restrict access to individuals, on the basis of facial recognition technology or other forms of AI. Weaponized drones equipped with Tasers, pepper spray and rubber bullets are already being used by police for crowd control, and could operate autonomously in the future using AI. Robots may in the future be able to physically restrain individuals. Actuator-equipped AI may also restrict freedom in more subtle ways. If human actions are delegated to AI systems, for

[148] High-Level Expert Group on Artificial Intelligence, “Ethics Guidelines for Trustworthy AI,” *European Commission*, July 4, 2019. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai; Santoni de Sio, Filippo, and Jeroen Van den Hoven, "Meaningful human control over autonomous systems: a philosophical account," *Frontiers in Robotics and AI,* Vol. 5, No. 15, 2018, DOI: 10.3389/frobt.2018.00015.
[149] Bozdag, Engin, and Jeroen Van Den Hoven, “Breaking the Filter Bubble: Democracy and Design,” *Ethics and Information Technology,* Vol. 17, No. 4, 2015, pp. 249–265.; Helberger, Natali, Kari Karppinen, and Lucia D’acunto, “Exposure diversity as a design principle for recommender systems,” *Information, Communication & Society*, Vol. 21, No. 2, 2016, pp. 191-207. DOI: 10.1080/1369118X.2016.1271900.

example in self-driving cars, it means that these actions are eliminated for the user of that technology: the driver will not be able to drive him- or herself, unless the self-driving vehicle includes a driving option for human drivers. Even more so, the scope of the action and the manner in which it is performed may be further restricted by the AI system. For example, the AI system may prohibit the vehicle from going off-road, thus limiting the freedom of the user to drive off-road.

The second general way in which we mentioned that AI systems can restrict freedoms is by providing information to third parties that can help them impose constraints on freedoms of individuals. These can be legitimate constraints, for instance constraints imposed by law enforcement in democratic countries, but they can also be constraints that violate fundamental rights, used by illiberal or authoritarian governments, by criminals, or by other private parties who fail to respect rights. Most importantly, AI can support this process through surveillance, profiling and data mining. Illiberal governments can use these processes to control populations. They can use AI to identify and track individuals and groups, to build up complex profiles of them using a variety of data sources, and to derive recommended actions to be taken to exercise control. They can use their monopoly on force to restrict freedoms and implement coercive actions. An example of this is the Chinese social credit system, which uses AI-driven mass surveillance to collect data about citizens and their behaviours, which is used to determine social credit scores that are then used, amongst others, to impose travel bans, limit unauthorized religious practices, and limit access to governmental services.

The risks of freedoms being limited are strongest with government use of AI, because governments have a monopoly on law enforcement and the use of force. However, private agents can also use surveillance and profiling to limit freedoms. Businesses, for example, already use it for surveillance on employees, and the information that employers gain can be used by them to limit freedom of speech, of assembly, or of movement. It is a question for debate when such limitations are justified and when they go too far.

Privacy

The increasing use and sophistication of AI technologies raises significant issues in relation to privacy. By and large, these issues ultimately stem from the fact that AI technologies often use as input data that is voluminous, from disparate sources, and about (groups of) individuals, and that it can generate profound, detailed and accurate insights on the basis of that data. While sometimes seen as a way to protect a specific private realm, privacy is often regarded as a means to realise other perhaps more fundamental values, such as freedom, autonomy, democracy, security, trust and friendship. It is a right provided by a number of international treaties on human rights.[150] One of the most serious potential consequences of a general lack of perceived privacy is that this may lead to a so-called "chilling effect" in society: a decrease in the legitimate exercise of civil liberties and rights (e.g., freedom of assembly, freedom of expression) that results from the fear among individuals of being watched.[151] There are different ways of conceptionalising privacy,[152] and there exist various types of privacy.[153] In the

[150] These include the International Covenant on Civil and Political Rights, and the European Convention on Human Rights. Also, in the EU, the Charter of Fundamental Rights of the EU provides for this right (and the right to data protection - articles 7 and 8). The EU's General Data Protection Regulation operationalises the right to data protection provided by the Charter of Fundamental Rights of the EU (art 8).

[151] Clarke, Roger, *Introduction to Dataveillance and Information Privacy, and Definitions of Terms*, 1997. http://www.rogerclarke.com/DV/Intro.htm

[152] DeCew, Judith, "Privacy," In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Spring 2018 Edition)*, 2018. https://plato.stanford.edu/archives/spr2018/entries/privacy

[153] Finn, Rachel, David Wright, and Michael Friedewald, "Seven Types of Privacy," In S. Gutwirth et al. (Eds.), *European Data Protection: Coming of Age*, Dordrecht: Springer, 2013.

remainder of this subsection, we describe how AI technology can harm the informational privacy of individuals in terms of their personal data and imagery, their personal communication, their behaviour and location, their thoughts and feelings, and their associations with other people.

First of all, many of the most important AI techniques utilise and produce large data sets, a fact that by itself increases the risks to privacy. The ability of these techniques to efficiently and effectively process large quantities of data might motivate their use, but their deployment might also necessitate the use of such data (e.g., for their proper functioning, many machine learning algorithms rely on being fed large volumes of data). In many applications, those data are bound to feature personal data, which may be mined and processed for any number of reasons (e.g., marketing opportunities, purchasing recommendations). The processing of larger data sets containing personal data, may then increase privacy and data protection risks simply by involving larger numbers of data subjects and more detailed personal data per subject.

Second, AI technologies possess unique capabilities in terms of identifying, monitoring and tracking individuals. They allow for the identification of people through speech, text, imagery, and web browsing data, amongst many other kinds of data. They can be used to monitor and track people's movements, with precision and in real-time, across different environments (e.g., in the home, at work, and in public spaces), across different devices (e.g., home speakers, smart appliances, mobile phones), and for large numbers of people at a time. Perhaps most worryingly, AI technology can in some cases be deployed to de-anonymise personal data that had been deemed anonymised.[154] Extensive identification, monitoring and tracking activities through AI systems in private and public spaces may diminish individuals' privacy of data and image, their privacy of location and space, their privacy of behaviour, their privacy of communication, and their privacy of association. Depending on the specifics of any given situation, this may in turn have negative effects on such values as freedom (e.g., freedom of speech, political freedom, freedom of association), autonomy, democracy, and security, and trust. Additionally, privacy harms through AI-based identification, monitoring and tracking may also increase the occurrence of errors (misidentifications) and unfair (e.g., discriminatory or biased) outcomes, depending on the use contexts, accuracy and potential biases of the algorithms. Furthermore, such privacy harms may contribute, over time, to shifting privacy norms, gradually lowering the expectations of anonymity in public spaces and other contexts.

Third, and related to the previous point, AI technologies enable sophisticated *profiling* and other predictive practices using data sets containing personal data. Profiling refers to the process of using pattern recognition and correlations to create *user profiles* that identify or represent people, and applying those profiles to analyse new data. To a large extent, profiling concerns the application of *group* profiles to individuals, which enables targeted servicing, refined price-discrimination and credit scoring, and identification of security threats.[155] By using patterns and correlations in data to make inferences, AI-based profiling and prediction permit far-reaching identification and monitoring of people's preferences and behaviours, even while seemingly trivial and/or anonymous data are used.[156] Such inferences can reveal highly sensitive information about individuals that these individuals may wish to suppress, and may not even be aware of themselves (e.g., predictions about their future

[154] Privacy International, "Artificial Intelligence", *Privacy International*, n.d. https://privacyinternational.org/topics/artificial-intelligence

[155] Hildebrandt, Mireille, and Serge Gutwirth, *Profiling the European Citizen. Cross Disciplinary Perspectives*, Dordrecht: Springer, 2008.

[156] Hildebrandt, Mireille, "Chapter 14: Who is Profiling Who? Invisible Visibility." In Gutwirth et al. (Eds.), *Reinventing Data Protection?*, Springer, 2009, pp.239-252.

health).[157] Dubious applications have already emerged or are emerging, including the assessment of individuals' emotions based on video, images, speech or text,[158] the identification of political leanings of a neighbourhood's residents based on the cars on the streets,[159] the prediction of a wide variety of physical and mental health conditions,[160,161,162] and the prediction of an individual's sexual orientation using facial imagery.[163] AI-based profiling and prediction thus poses significant risks to people's privacy of data and image, their privacy of behaviour, and their privacy of thought and feelings, amongst others. Violations of these types of privacy could lead to harms to freedom, autonomy, democracy, and security, and trust, as well as bias, discrimination, manipulation, errors, and an overall decrease in wellbeing.

Fourth, AI technology can generally be used to gather and analyse personal data in a highly inconspicuous manner. Oftentimes, humans have no way of knowing that behind a simple security camera there is AI system that tracks their every move, or that they are being profiled on the basis of their online behaviour. This lack of knowledge means that it is generally hard to guard oneself against any AI-based privacy intrusions, and it may contribute to the aforementioned "chilling effect" on society.

Finally, it is worth noting that while any significant harms to privacy as a result of the use of AI systems can be considered highly problematic, it is also the case that many of the most prominent and promising AI techniques today (i.e., most importantly, the machine learning algorithms that rely on vast quantities data) will offer worse performance if significant privacy and data protections are put in place. From a consequentialist perspective, privacy and data protection may thus need to be balanced against the positive effects of making full use of such techniques, which can include economic growth, improved health, et cetera.

Justice and fairness

This section concerns how AI may impact just and fair processes in society. In particular, it will focus on distributive justice: the socially just allocation of goods in society. This issue is central in most theories of justice. They hold that justice is, to a large extent, about the fair distribution of social goods in society. Social goods are goods that that are not basic mental and bodily abilities like health, strength, and intelligence, but goods that can be allocated and distributed in society, such as income, rights, housing, and means of transportation. Theorists of justice, like John Rawls (1971), are especially concerned with primary social goods, which are social goods that every rational human being is

[157] Ibid.

[158] See for instance, IBM's Watson Tone Analyser (https://www.ibm.com/watson/services/tone-analyzer/), Microsoft's Azure products (https://azure.microsoft.com/en-gb/services/cognitive-services/face/), and Affectiva (https://www.affectiva.com/).

[159] Myers, Andrew, "An artificial intelligence algorithm developed by Stanford researchers can determine a neighborhood's political leanings by its cars," *Stanford News*, November 28, 2017. https://news.stanford.edu/2017/11/28/neighborhoods-cars-indicate-political-leanings/

[160] Rajkomar, Alvin, "Scalable and accurate deep learning with electronic health records," *NJP Digital Magazine*, 2018. https://www.nature.com/articles/s41746-018-0029-1

[161] IBM, "With AI, our words will be a window into our mental health", *IBM*, n.d. https://www.research.ibm.com/5-in-5/mental-health/

[162] Burgess, Matt, "Now DeepMind's AI can spot eye disease just as well as your doctor", *Wired*, August 13, 2018. https://www.wired.co.uk/article/deepmind-moorfields-ai-eye-nhs

[163] PERVADE: Pervasive Data Ethics, ""The study has been approved by the IRB": Gayface AI, research hype and the pervasive data ethics gap," *Medium (PERVADE: Pervasive Data Ethics)*, November 30, 2018. https://medium.com/@pervade_team/the-study-has-been-approved-by-the-irb-gayface-ai-research-hype-and-the-pervasive-data-ethics-3b36c5a53eec

presumed to want because they have a use whatever a person's plan of life.[164] Rawls has argued that the primary social goods include rights. liberties and opportunities (including freedom of thought, freedom of association, freedom of movement, free choice of occupation, equal opportunities in careers, political liberties and the rights and liberties covered by the rule of law), income and wealth, as well as the social bases of self-respect. Van den Hoven and Rooksby (2008) have recently argued that information has nowadays also become a primary social good, because good access to information and information technologies has become vital for proper functioning in society.[165]

Different theories of justice hold different positions on which distributions of primary social goods in society are fair and which ones are unfair. In general, however, theories of justice agree that in principle people should have equal rights, liberties and opportunities. This includes, amongst others, equal treatment under the law, equal opportunity in hiring, and equal access to certain social goods and services (e.g., education, healthcare, utilities, social services). It also requires the absence of legally enforced social class boundaries and the absence of discrimination based on inalienable parts of one’s identity, including gender, race, age, sexual orientation, national origin, religion, income, property, health, disability and opinions. Discrimination is the unequal treatment of individuals or groups based on such characteristics, in a way that denies them opportunity or treats them worse than others, solely because of these characteristics.

How can the use of AI systems result in unfair treatment, or, conversely, how can it restore fairness? In this section, we will focus on the first question, while also paying some attention to the second. First, let us consider four ways in which AI systems can contribute to, or be involved in, the unfair treatment of individuals and groups:

(1) *Inequality of access to AI systems and services.* AI systems can confer considerable benefits to their users, providing them with useful information and services. For some groups, however, access to these systems, or to some of its functional features, can be limited. For example, people may not have access because of financial constraints. Others may not have access because the system’s interface requires too much computer literacy, or because it has not been designed to support visual, cognitive or other handicaps. Also, certain programs or databases may have their access restricted to certain organisations or groups without a strong justification. Inequality of access to AI systems and services that provide vital (i.e., primary) social goods therefore raises questions of fairness and justice.

(2) *Functional bias.* This pertains to AI systems offering functionality that serves the interest of certain social groups of users, while less so those of others. Any technological artefact presupposes a particular user with certain interests and goals. For example, AI-powered financial planning software may presuppose that users have a secure job, and that they have an investment portfolio, while in fact there are users that do not meet these criteria, and therefore the software is less useful for them. This is functional bias: a bias in the functional features of AI systems that serves the needs and interests of certain individuals and social groups better than those of others. In case there are no alternative systems available that do serve these underrepresented individual and groups, one might say that functional bias results in unfair outcomes, as some groups are better served than others.

[164] Rawls, J. (1971), A Theory of Justice, Cambridge, MA, Harvard University Press.

[165] Van den Hoven, J., & Rooksby, E. (2008). Distributive Justice and the Value of Information: A (Broadly) Rawlsian Approach. In J. van den Hoven, & J. Weckert (eds.), *Information Technology and Moral Philosophy*, Cambridge: Cambridge University Press, 376-396.

(3) *Algorithmic bias.* Bias in treatment of individuals and social groups represented by the system or otherwise affected by the system's decisions or recommendations. This is called algorithmic bias.

(4) *Unfairness in social effects of AI.* Even if no functional and algorithmic bias were to be present in an AI, and universal access were secured, there could still be social effects of the use of AI system that are unfair by resulting in unjust distributions of primary social goods. E.g., unemployment, power asymmetries. The use of AI by particular groups or organisations causes other groups to be treated unfairly or have less opportunity, even if this was not the intention of the system's design.

Out of these four types of unfairness associated with AI, by far the most attention has gone to the issue of algorithmic bias. Let us now discuss this issue in more detail.

Algorithmic fairness concerns whether an AI system makes decisions and produces results that do not unjustly discriminate against groups or individuals.[166] The potential biases that may undermine an AI's fairness are *input data bias*, *computational bias*, or *outcome bias*.[167]

Input data biases are implicit or explicit distortions within the data an AI analyses. As machine learning systems develop and refine their algorithms by analysing training data, they are susceptible to reinforcing any implicit or explicit biases contained within that data. For example, the IJB-A dataset developed by the US National Institute of Standards and Technology (NIST) in 2015 for testing facial recognition systems was found to significantly overrepresent lighter-skinned faces.[168] AIs trained to perform facial recognition based on this data would therefore be more accurate at recognising the faces of people with lighter skin colours. Similarly, analysis of large text datasets has uncovered implicit gender bias within large datasets of news reports.[169] Occupations such as 'homemaker', 'nurse', and 'librarian' were found to be strongly correlated with the female gender, while occupations such as 'architect', 'philosopher', and 'financier' were strongly correlated with males.[170] As a result, an AI trained using similar data would associate these roles with a particular gender. In another case, Google was showing men advertisements for higher paying jobs, while women were shown more generic advertisements.[171]

Removing details such as race or gender from training data does not necessarily mean that training an AI trained on that data will produce unbiased results. The AI may still become biased by drawing conclusions on details that serve as proxies for the deliberately omitted information. This is possible as other details may have a strong correlation with omitted details such as race or gender.[172] For example, the geographic location of someone's residual address may serve as a proxy value for race

[166] Springer, A., Garcia-Gathright, J., and Cramer, H., 'Assessing and Addressing Algorithmic Bias – But Before We Get There', *2018 AAAI Spring Symposium Series*, March 2018, pp. 450-454. https://www.aaai.org/ocs/index.php/SSS/SSS18/paper/viewPaper/17542.

[167] Ibid., p. 451.

[168] Buolamwini, J., and Gebru, T., 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification', Proceedings of the 1st Conference on Fairness, Accountability, and Transparency, PMLR, Vol. 81, pp. 77-91, 2018.

[169] Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A., 'Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings', 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, 2016. http://papers.nips.cc/paper/6227-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.

[170] Ibid.

[171] Datta, A., Tschantz M. C. & Datta, A. (2015). Automated Experiments on Ad Privacy Settings A Tale of Opacity, Choice, and Discrimination. *Proceedings on Privacy Enhancing Technologies*, 2015 (1), 92–112.

[172] O'Neil, *Weapons of Math Destruction*, London, Allen Lane, pp. 23-27, 2016.

or socio-economic status.[173] An instance of this can be seen in Amazon's "prime-lining", where low-income minority neighbourhoods were excluded from their service. In this case, the "low income" and the "minority" labels were actually proxies for race.[174]

Computational biases are the result of choices the developers make in creating and refining an AI.[175] They may emerge from the developers' understanding of the values of the users and stakeholders affected by the AI.[176] In creating an AI, the developers must create an abstract model that represents the actual phenomena or population that the system is intended to evaluate. This model is also limited by the technical constraints of the system and the available data. As a result, the developed model will be an incomplete representation of the actual phenomena or population.[177] The developers therefore must decide what data best represents what their system is intended to model, and what aspects of the phenomena or population can be simplified or omitted without compromising the accuracy of the system's decisions. These design decisions may be affected by the developers' implicit or explicit biases, and so may affect the fairness of the AI's decisions.

The developers' decisions about the acceptable level of accuracy may also introduce unfairness into an AI. The AI's output may also be biased by the developers' decisions about how to optimise the system's accuracy.[178] For example, an AI is likely to make both false positive and false negative responses about input data. The rate at which these errors are made can be modified by changing parameters within the system. Depending on the context, false positives and false negatives may have significant differences in their acceptability: false positives in a recommendation system for books and movies is less significant than a false negative in a medical screening system. Improving overall accuracy may also cause the system to produce different levels of false positives and false negatives for a certain group, making them the target of more decision errors than other groups.[179]

Outcome biases may arise from the feedback mechanisms that exist between the AI and the environment affected by its decisions.[180] For example, an AI that predicts criminal activity in urban areas will allocate more police to areas determined to have a high likelihood of crime than to others. The increased police presence will result in crimes being reported that would have otherwise gone unnoticed. As a result, there will be more reported crimes in that area, which reinforces the AI's decisions that crime is likely to occur in that area, even if other areas may have equivalent or greater levels of unreported crime.[181]

Outcome biases may lead to existing social inequalities becoming further entrenched within society. AI systems may reinforce inequalities by allowing people to be targeted with information that reinforces their prejudices, or with offers that exploit their vulnerabilities. The social media posts of

[173] Veale, M., and Binns, R., 'Fairer Machine Learning in the Real World: Mitigating Discrimination Without Collecting Sensitive Data', *Big Data & Society*, Vol. 4, No. 2, July-December 2017.
[174] Johnson, J. A. (2018). Open Data, Big Data, and Just Data. In *Toward Information Justice*, ed. by J. A. Johnson, 23-49.
[175] Springer, A., Garcia-Gathright, J., and Cramer, H., 'Assessing and Addressing Algorithmic Bias – But Before We Get There', *2018 AAAI Spring Symposium Series*, March 2018, pp. 450-454. https://www.aaai.org/ocs/index.php/SSS/SSS18/paper/viewPaper/17542., p. 451.
[176] Dobbe, R., Dean, S., Gilbert, T., and Kohli, N., 'A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics', 2018 Workshop on Fairness, Accountability and Transparency in Machine Learning during ICML 2018, Stockholm, Sweden, 2018. https://arxiv.org/abs/1807.00553.
[177] Ibid.
[178] Ibid.
[179] Ibid.
[180] Ibid.
[181] O'Neil, C., *Weapons of Math Destruction*, London, Allen Lane, pp. 84-91, 2016.

individuals may be used by machine learning systems to determine what political messages will be the most effective in persuading them to support a political candidate, party, or policy.[182] Companies may use machine learning to target advertising towards individuals more likely to respond to their messages. While this may be to the individual's benefit by offering them products and services that are specific to their needs, it may also be used to exploit the individual's vulnerabilities. It may further inequalities by allowing advertisers to target vulnerable people with predatory advertising for products and services they cannot afford.[183]

As described above, AI systems that are developed using biased input data are likely to reflect those inequalities in their decisions. For example, an AI that estimate the risks of a criminal re-offending after release might unfairly classify people of a certain race as being more likely to re-offend due to it being trained using data that was itself biased by discriminatory police activity that unfairly targeted people of that race.[184]

Let us now turn to the *sources* of bias and potential ways to mitigate of bias. Some of the forms of unfairness associated with AI can be traced back to the composition of the design teams of this technology and their beliefs and prejudices. Many AI developers are affluent white men, who may introduce unconscious biases based on their lived experience into the systems they create.[185] For example, women make up only 10% and 15% of AI researchers at Google and Facebook, respectively.[186] There are also low levels of racial diversity in major AI companies.[187] This 'diversity crisis' in AI has wide-ranging effects, both within companies and organisations who create AI systems and the broader community affected by the products they create.[188] For example, the overrepresentation of lighter-skinned people in the facial recognition dataset mentioned earlier might not be readily apparent to developers working in a predominately white workforce.

Another reason for unfairness can be found in the functioning of the market. AI systems are not necessarily designed for fairness and the greater good, but rather they are, in most cases, designed to generate profit. Demand for AI system is highest for those who are already powerful and wealthy, and this fact generates a potential for unfair outcomes as those who are not powerful and wealthy could see their opportunities and liberties reduced as a result of this unequal distribution of AI capabilities.

Yet another cause of unfairness can be found in the absence of sufficient consideration of issues of fairness in the design and implementation of AI systems. Designers often do not recognize this as an issue that needs to be considered. Even if they do, they often have not been trained in ethical assessment and in methodologies for the inclusion of ethical considerations in design. In addition, benefits could be had from usability studies and user testing with a diverse user base, but often, such testing does not occur, or does not occur with a diverse enough user group.

[182] Berghel, H., 'Malice Domestic: The Cambridge Analytica Dystopia', *Computer*, Vol. 51, No. 5, pp. 84-89, May 2018.

[183] O'Neil, C., *Weapons of Math Destruction*, London, Allen Lane, pp. 84-91, 2016.

[184] Zerilli, J., Knott, A., Maclaurin, J., and Gavaghan, C., 'Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?', *Philosophy & Technology*, 2018, https://doi.org/10.1007/s13347-018-0330-6.

[185] Crawford, K., 'A.I.'s White Guy Problem', *The New York Times,* p. SR11, June 26, 2016. https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html.

[186] West, S. M., Whittaker, M., and Crawford, K., 'Discriminating Systems: Gender, Race, and Power in AI', 2019, p. 5. https://ainowinstitute.org/discriminatingsystems.html.

[187] Ibid.

[188] Ibid., p. 3.

A final potential cause of unfairness lies in the moral opacity of AI technology, even for designers of the technology. Designers often do not have good insight into the detailed operation of the system. This is especially the case for neural network AI and machine learning. As a result, it is difficult for them and others to recognize algorithmic biases that emerge in the workings of the technology.

Mitigation strategies for unfairness in AI can be related to these four potential causes. First of all, diversification of design teams of AI is a priority and will help mitigate bias. In addition, training personnel to recognize their own biases and prejudices and overcome them, particularly in relation to their design practice, could also contribute to the reduction of bias and unfairness. Third, governments and NGOs, as well as companies, can play a role in levelling the playing field and in ensuring that universal access to important AI systems and services is improved, that functional bias is reduced, and that further unfair effects of AI applications are mitigated. Fourth, training of designers in ethical reflection and ethics by design methodologies can help them better consider ethical issues in general, and fairness issues specifically, in design. Specific attention to algorithmic fairness methodologies will be of particular help here. In addition, it would be beneficial if industry engaged more often in user testing with a diverse user base. Fifth and finally, transparency of AI is a prerequisite for AI developers to adequately understand and diagnose biases in AI systems, and therefore the development of adequate transparency and explainability methods for AI is another way to help reduce bias and unfairness in AI.

To conclude our discussion on justice and fairness, let us briefly consider AI technology as a means for restoring fairness. AI has a potential for contributing to fairer practices, in the first place because it can compensate for biases in decisions that are normally taken by humans. AI systems that are well-designed to minimize bias and make decisions on objective grounds can potentially be fairer than humans in decision-making, discarding some of the biases and prejudices that exist in humans. Another way in which AI can make conditions fairer is by providing disadvantaged individuals and groups with powers and opportunities that they would not have had without AI, for example in producing better adaptive technologies for people with disabilities, or for inexpensive services or advice for people who cannot pay for human professionals to provide these services. Some companies and governments are supporting the notion of "AI for good" to develop AI systems and applications whose primary purpose is to help solve societal challenges, including challenges like reducing inequality, reducing poverty, and developing more just institutions.

Responsibility and accountability

In the ethics of AI and robotics literature, there has been a growing debate on the ethical implications of AI and robotics technology in terms of accountability and responsibility. As has been outlined in the preceding sections of this report, AI systems are increasingly being used to make decisions that can have very significant consequences for individuals, organisations and society at large. Given these risks, many have argued it should be possible to hold individuals and organisations accountable for the harms that result from the AI systems they use and/or develop. The term *algorithmic accountability* has been used to refer to (1) the assignment of responsibility for the development of an algorithm (and, by effect, any AI system) and any societal implications of its use, and (2) accountable systems that include compensatory mechanisms for any harm that may come to pass.[189] Accountability in this sense serves to ensure the responsible use and development of such systems by deterring illegal,

[189] Donovan, Joan, Robyn Caplan, Jeanna Matthews, "Algorithmic accountability: A primer. Data & Society Tech Algorithm Briefing: How Algorithms Perpetuate Racial Bias and Inequality," Washington, DC, USA, 2018. https://datasociety.net/wp-content/uploads/2018/04/Data_Society_Algorithmic_Accountability_Primer_FINAL.pdf

reckless or otherwise irresponsible behaviour by those using AI systems,[190] and by generating a "self-reflective feedback loop for citizens and society" that may expose entrenched biases and power relations.[191,192] There are, however, a number of issues that complicate accountability.

First of all, it is difficult to have algorithmic accountability when AI systems lack transparency.[193,194] For an AI system or algorithm to be transparent, its purpose, inputs and operations should be knowable by its stakeholders, so that they can understand how its decisions are arrived at. Many AI systems that are currently being developed lack transparency and can be characterised as "black boxes", where we can see input-output relations but we do not know how and why they are produced. As explained more thoroughly further on in this subsection (under "Transparency and explainability"), a lack of transparency can be due to: (1) the sheer complexity of an algorithm; (2) the inherently uninterpretable nature of an algorithm; or (3) the inability of lay persons to understand the explanations for an algorithm's workings. Neural networks, and especially deep learning algorithms, are some of most problematic algorithms, as these kinds of algorithms provide no way of explaining how they reach their results.[195] (Please note that a more thorough discussion of ethical issues in relation to transparency and explainability is provided further on in this subsection.)

The lack of transparency in AI systems makes it harder to ascribe responsibility to any individual(s) or organisation(s) for the proper functioning of such systems, and to hold them accountable for any harms these systems might cause. This is because it is typically only justified to assign responsibility and attribute blame for some harm-causing action when an actor had some degree of control over, and intentionality in, carrying out said action.[196] And in order have control over an action, one needs to have an understanding of what the action entails. In the case of an AI system, having control thus requires that the system's workings are transparent (from the subjective perspective of the individual who need to have control). When an AI system is completely opaque, such as in the case of a deep-learning-based system, developers and users have no control over it, as they cannot predict what it will do, and this will lead to what is called a "responsibility gap" where no one can be held responsible for the actions of the system.[197] In recent years, there have been diverging and contentious efforts to close or remediate this responsibility gap, which include approaches that consider machine-learning algorithms as mere tools for strictly human decision-making and action,[198] approaches that consider these systems as "functional" moral agents with (quasi) moral responsibility of their own,[199] and approaches that take a middle ground and distribute responsibility over a network of human and

[190] Koene, Ansgar, Chris Clifton, Yohko Hatada, Helena Webb, Menisha Patel, Caio Machado, Jack LaViolette, Rashida Richardson, and Dillon Reisman, "A governance framework for algorithmic accountability and transparency," *European Parliamentary Research Service*, 2019. http://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_STU(2019)624262

[191] Ibid.

[192] Tutt, Andrew, "An FDA for Algorithms," *SSRN Electronic Journal*, 2016.

[193] Ananny, Mike, and Kate Crawford, "Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability," *New Media & Society*, Vol. 20, No. 3, 2016, pp. 973–989.

[194] Diakopoulos, Nicholas, "Accountability in Algorithmic Decision-Making," *Queue*, Vol. 13, No. 9, 2015, pp. 126–149.

[195] Lipton, Z. C. (2017). The mythos of model interpretability. arXiv preprint arXiv:1606.03490.

[196] Matthias, Andreas, "The responsibility gap: Ascribing responsibility for the actions of learning automata," *Ethics and Information Technology*, Vol. 6, No. 3, 2004, pp. 175–183.

[197] Ibid.

[198] E.g., Bryson, J. J. (2010). Robots should be slaves. In Y. Wilks (Ed.), Close engagements with artificial companions: Key social, psychological, ethical and design issues (pp. 63–74). Amsterdam: John Benjamins.

[199] E.g., Stahl, B. C. (2006). Responsible computers? A case for ascribing quasi-responsibility to computers independent of personhood or agency. Ethics and Information Technology, 8(4), 205–213.

machine components.[200,201] As yet, at least in the philosophical literature, there seems to be no consensus, however, on how to properly deal with the responsibility gap and other problems with responsibility ascription.[202]

A lack of transparency in AI systems may further complicate algorithmic accountability as it often indicates that there is no mechanism to correct and improve decision-making procedures that are considered erroneous or unfair. In addition, a lack of transparency makes it difficult to notice when harms have occurred in the first place—harms that could justify compensatory action.

While a lack of transparency in AI systems may be ethically highly problematic in term of its effect on accountability, having a lot of it, it has been argued, may also come at a cost. As is explained in more detail further on in this subsection (under "Transparency and explainability"), requirements for accountability and transparency that are too stringent could lead to unnecessary costs and stifled innovation. Ultimately, a sensible balance may have to be struck between the need for transparency and accountability, on the one hand, and economic interests relating to system performance and the protection of intellectual property rights, on the other.

A final set of challenges for algorithmic accountability lie at the level of governance. To ensure that developers and users take their responsibilities in terms developing and using AI systems in responsible ways, many argue for new laws, government policies, ethical guidelines, and industry self-regulation initiatives. For the most part, efforts in these areas are currently only in the beginning stages. State intervention through, for example, taxes and subsidies can promote better algorithmic behaviour.[203] Furthermore, ethical guidelines, it is argued, need to emphasise the need for more transparency and explainability, and need to be clear on how to ascribe responsibility, and who to hold accountable when things go wrong. Finally, algorithms need to be auditable by independent organisations to ensure public accountability. As yet, such independent auditing is rarely used since there are no widely accepted industry standards and guidelines for assessing social impact.[204]

Safety and security

While the potential threat posed by human-level or super-intelligent AI systems is frequently discussed,[205] there are already pressing concerns about the safety and security of current AI applications. Safe and secure AI requires reliable systems that justify the trust placed in them by ensuring that they make appropriate decisions, and that they are resilient to system faults and deliberate attacks.

Trust may be fostered by clear accountability for errors or faults occurring within an AI. However, the methods used within many AI systems make it difficult (if not impossible) to identify what caused the

[200] E.g., Hanson, F. A. (2009). Beyond the skin bag: On the moral responsibility of extended agencies. Ethics and Information Technology, 11(1), 91–99.

[201] Gunkel, David, "Mind the gap: responsible robotics and the problem of responsibility," *Ethics of Information Technology*, 2017.

[202] Other factors that may complicate responsibility ascription include the involvement of a large number of people in building an advanced AI system through what may be a complex development process ("the problem of many hands"), and decision-making processes where AI systems are helping experts to make decisions (who is responsible is ultimately responsible for the effects of the decisions that are made?). Friedman, B., 1990. "Moral Responsibility and Computer Technology," Paper Presented at the Annual Meeting of the American Educational Research Association, Boston, Massachusetts.

[203] Donovan, 2018, op. cit.

[204] Ibid.

[205] Müller, V. C. (ed.), *Risks of Artificial Intelligence*, Boca Raton, CRC Press, 2016; Bostrom, N., *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, 2014.

error. This is the 'black box' problem of AI.[206] An AI may be opaque in three ways: intentional secrecy by the developers or operators, technical illiteracy, and the mathematical optimisation created by the AI's training that comes at the cost of being understandable to human interpretation.[207] The lack of transparency in how AI systems operate is a serious obstacle towards establishing clear lines of accountability and therefore making them trustworthy systems. The current research interest in Explainable AI (XAI) systems that disclose the hidden reasoning used in their own or another system's decisions is one response to this problem.[208]

Like any computer system, hardware faults that may cause errors in an AI's decision making.[209] An AI may also be subject to deliberate attacks. For example, the integrity of machine learning systems may be undermined through manipulating how the system collects and processes data, by altering the model it develops through analysis of the training data, or by manipulating the system's output.[210] Microsoft's Twitter chatbot Tay is an example of how an AI can be manipulated through directing biased input towards it. The chatbot was designed to learn human figures of speech through interaction with human users on Twitter; however, Microsoft was forced to take it offline within 24 hours after mischievous users had trained it to produce messages containing sexist, racist, and anti-Semitic language.[211]

Since many AI systems develop their internal models of how to respond to input from analysing training data, omissions or ambiguities in the training data may lead to unexpected results when faced with actual data. For AI incorporated into cyber-physical systems, such as automated vehicles, the wrong response to ambiguous data may have fatal consequences. For example, in 2016 a Tesla Model S car operating in 'Autopilot' mode confused a white truck for a clear sky, and as a result caused an accident resulting in the driver's death.[212] While there were clear warnings that the driver should remain ready to override the car's behaviour, this example also demonstrates the possibility of *overtrusting* an AI system.[213] Ensuring that an AI is safe to use requires its users to have the appropriate level of trust in the AI's capabilities. As discussed in section 5.1.1, an awareness of the risks of relying on an AI system is necessary to ensure that it is used safely and appropriately. Too much trust encourages an uncritical acceptance of the AI's outputs, with potentially harmful consequences. Conversely, too little trust in an AI means that the safety benefits of automation are lost.[214]

[206] Castelvecchi, D., 'The Black Box of AI', *Nature*, Vol. 538, No. 7623, pp. 20-23, 6 October 2016.

[207] Burrell, J., 'How the Machine 'Thinks': Understanding Opacity in Machin Learning Algorithms'*, Big Data & Society*, Vol. 3, No. 1, June 2016.

[208] Miller, T., 'Explanation in Artificial Intelligence: Insights From the Social Sciences', *Artificial Intelligence*, Vol. 267, pp. 1-38, February 2019.

[209] Hanif, M. A., Khalid, F., Putra, R., Rehman, S., and Shafique, M., 'Robust Machine Learning Systems: Reliability and Security for Deep Neural Networks', 2018 IEEE 24th International Symposium on On-Line Testing And Robust System Design (IOLTS), 2-4 July 2018.

[210] Papernot, N., McDaniel, P., Sinha, A., and Wellman, M., 'Towards the Science of Security and Privacy in Machine Learning', p. 4, 2016. https://arxiv.org/abs/1611.03814.

[211] Wolf, M. J., Miller, K., and Grodzinsky, F. S., 'Why We Should Have Seen That Coming: Comments on Microsoft's Tay "Experiment", and Wider Implications", *ACM SIGCAS Computers and Society*, Vol. 47, No. 3, pp. 54-64, September 2017.

[212] Tesla, 'A Tragic Loss', June 30, 2016. https://www.tesla.com/blog/tragic-loss.

[213] Wagner, A. R., Borenstein, J., and Howard, A., 'Overtrust in the Robotic Age', *Communications of the ACM,* Vol. 61, No. 9, pp. 22-24, 2018.

[214] Lee, J. D., and See, K. A., 'Trust in Automation: Designing for Appropriate Reliance', *Human Factors*, Vol. 46, No. 1, pp. 50-80, Spring 2004.

Dual use and misuse

Dual use technologies have both beneficial uses as well as the potential to cause significant harm if used maliciously.[215] AI is such a technology as it may be abused to create new threats or to make existing harms easier to perform through automation.[216] Such malicious uses may be directed to undermine *digital* security towards automating aspects of network intrusions, threaten *physical* security through interference with cyber-physical systems (such as automated vehicles), or damage *political* security by undermining trust in political leaders and institutions.[217] These abuses may be performed by individuals, groups, or governments.

Two examples of how AI may be used to undermine digital security is by using it to create unique malware (malicious software) and social engineering attacks. AI may be used to develop malware tailored to be undetectable to existing malware detection systems.[218] It may also be used to enhance *spear-phishing* attacks, where personalised fraudulent messages are sent to specific individuals to mislead them into sharing information with the attacker or performing some action for the attacker's benefit.[219] An AI could be used to automate the extensive research required to create a convincing message that would avoid causing suspicion. Similarly, the automation AI makes possible might allow for greater numbers of individuals to be targeted with convincing fraudulent messages.

While spear-phishing is an example of how AI may automate an existing threat, it may also be used to create new security threats. The ability to train an AI to create new output based on an existing dataset creates the possibility of using it to create material that impersonates the works, image, or voice of another person. This creates new opportunities to spread disinformation for political gain by producing convincing video and audio recordings of leaders.[220] AI-assisted disinformation campaigns could also be used to create far greater amounts of propaganda material in a shorter time than was previously possible. For example, in February 2019 the research organisation OpenAI refused to publicly release its GPT-2 machine learning system due to concerns that it could be used to create large numbers of convincing false news stories.[221]

A similar abusive use of AI is to create so-called 'deepfakes': images or video footage modified using an AI to create misleading, malicious or humiliating depictions of people.[222] While tampering with

[215] Selgelid, M., 'Dual-Use Research Codes of Conduct: Lessons from the Life Sciences', *NanoEthics*, Vol. 3, No. 3, pp. 175-183, 2009.

[216] Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., Ó hÉigeartaigh, S., Beard, S., Belfield, H., Farquhar, S., Lyle, C., Crootof, R., Evans, O., Page, M., Bryson, J., Yampolskiy, R., and Amodei, D., 'The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation', 2018. https://www.eff.org/files/2018/02/20/malicious_ai_report_final.pdf.

[217] Ibid.

[218] King, T. C., Aggarwal, N., Taddeo, M., and Floridi, L., 'Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions', *Science and Engineering Ethics*, 2019. https://doi.org/10.1007/s11948-018-00081-0.

[219] Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., Ó hÉigeartaigh, S., Beard, S., Belfield, H., Farquhar, S., Lyle, C., Crootof, R., Evans, O., Page, M., Bryson, J., Yampolskiy, R., and Amodei, D., 'The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation', p. 18, 2018. https://www.eff.org/files/2018/02/20/malicious_ai_report_final.pdf.

[220] Ibid., p. 20.

[221] Radford, A., Wu, J., Amodei, D., Amodei, D., Clark, J., Brundage, M., and Sutskever, I., 'Better Language Models and Their Implications', OpenAI, February 14, 2019. https://openai.com/blog/better-language-models/.

[222] Chesney, B., and Citron, D., 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security', SSRN, 2018. https://ssrn.com/abstract=3213954

photographs, audio recordings, and video footage for malicious purposes is nothing new, AI allows for more sophisticated manipulation that is more difficult to identify as fraudulent. This form of misuse is particularly concerning due to its relative accessibility: the term ‘deepfakes’ itself comes from communities of Internet forum users who used machine learning systems to place celebrities’ faces on people shown in pornographic videos.[223] As this technology becomes even more accessible, the capability to create convincing depictions of anyone could be easily abused in personal disputes to impersonate or humiliate others.

Mass unemployment

Automation by means of artificial intelligence and robotics will inevitably lead to job losses in the future. There is consensus that many jobs could be affected. However, wildly different estimates exist of how many jobs will be lost, and of how many new jobs will be created due to the introduction of AI and robotics. The estimates vary from half of all jobs being lost to no net loss of jobs or even job growth. On one end of the spectrum, a much-cited report by Frey & Osborne (2013) claims that 47% of jobs in the United States are at risk because of automation.[224] On the other end of the spectrum, a recent editorial in Skynet Today cites studies to argue that there will only be modest job loss due to AI, which will be more than offset by new jobs.[225]

An OECD-commissioned study across 32 countries finds that about 14% of jobs in OECD countries could be lost because of automation because they are highly automatable (automation probability of over 70%). In addition, another 32% of jobs have a risk of between 50 – 70% to be automated.[226] This means that a total of 46% of jobs are at a high risk of being automated, dovetailing the Frey & Osborne study. The study also points out that new jobs will be created, but finds it hard to determine which ones and how many. A McKinsey report investigates possible displacement of jobs across 46 countries, and finds that an average of 15% could be displaced by 2030, but cautions that the bandwidth of their estimates ranges from almost zero to thirty percent.[227] A PriceWaterhouseCoopers study estimates that by the 2030s, the percentage of jobs that could be automated ranges between 22% and 44% for different countries.[228]

Regarding the types of jobs that would be lost, studies tend to agree that low- and middle-skilled jobs are most at risk, and high-skilled jobs are at low risk. Most at risk are white-collar and blue-collar jobs that are routine-based. White-collar jobs of this type include clerical workers such as data entry clerks, accounting and payroll clerks, and secretaries, as well as auditors, bank tellers, cashiers, sales workers, and financial analysts. Blue-collar jobs include jobs in transportation and storage, manufacturing, and

[223] Ibid.
[224] Frey, Carl Benedikt, and Michael A. Osborne, "The future of employment: How susceptible are jobs to computerisation?," *Technological forecasting and social change,* Vol. 114, 2017, pp. 254-280. DOI: 10.1016/j.techfore.2016.08.019.
[225] Job loss due to AI — How bad is it going to be? (2019, February 4). *Skynet Today.* Retrieved at: https://www.skynettoday.com/editorials/ai-automation-job-loss.
[226] Nedelkoska, Ljubica, and Glenda Quintini, "Automation, skills use and training," *OECD Social, Employment and Migration Working Papers*, No. 202, OECD Publishing, Paris, 2018, http://dx.doi.org/10.1787/2e2f4eea-en.
[227] Manyika, James, Susan Lund, Michael Chui, Jacques Bughin, Jonathan Woetzel, Parul Batra, Ryan Ko, and Saurabh Sanghvi, “Jobs Lost, Jobs Gained: Workforce Transitions in a Time of Automation,” *McKinsey Global Institute*, 2017.
[228] Hawksworth, John, Richard Berriman, and Saloni Goel, "Will robots really steal our jobs? An international analysis of the potential long term impact of automation," *PricewaterhouseCoopers LLP,* Retrieved at https://www.pwc.com/hu/hu/kiadvanyok/assets/pdf/impact_of_automation_on_jobs.pdf.

construction.[229] Less at risk are high-skilled jobs, according to most studies, and low- and medium-skilled jobs that are not easy to automate because they involve non-routine tasks or take place in unpredictable environments. These include both low-skilled and medium-skilled jobs, many of them in in education, healthcare, some of the so-called pink-collar jobs in the services sector, as well as blue-collar professions like gardener and plumber.

The McKinsey report claims that advanced economies will be more affected by automation than developing ones, because the higher wage rates in these countries provide bigger economic incentives to automate. [230] There is disagreement, however, whether the impact of automation will be greater for low-skilled or medium-skilled workers. Muro, Maxim and Whiton (2019) argue that in recent decades, automation has led to job and income losses for middle-waged and middle-skilled people, but that there has been both job and wage growth for low-wage and high-wage workers in the same period, and they expect that this will continue for automation fuelled by AI.[231] However, Nedelkoska and Quintini (2018) find that low-skilled jobs are on the whole more at risk than medium-skilled jobs for automation, a result that is mirrored in other studies as well. [232]

Studies do not agree on the impact of automation along gender and ethnic lines. Regarding gender, Muro, Maxim and Whiton (2019) find in their study of US employment that men are more at risk to lose their job due to automation than women, 43% to 40%, due to their overrepresentation in manufacturing, transportation and construction jobs that are at risk for automation, and due to the overrepresentation of women occupations in sector like health care, personal services, and education that are relatively safe.[233] World Economic Forum (2018b) has found that 57% of jobs at risk for disruption belong to women.[234] They take into account that, according to their analysis, at-risk jobs in professions dominated by men have more reskilling and job transition options than those in professions dominated by women. Regarding ethnicity, Muro, Maxim and Whiton find that in the United States, Hispanic and Black workers are more at risk than white workers (47% and 44% vs. 40%), and Asian workers are less at risk (39%).[235]

Next to jobs being lost due to AI-driven automation, studies suggest that many jobs will be transformed as well. Routine tasks in them will be eliminated, and people will have to retrain and upgrade their skills to retain a competitive advantage and to work with the new technology. Many wages are moreover likely to fall, since AI-driven automation makes labour less profitable, especially low- and middle-skilled labour. In addition, new jobs will be created, including a relatively small number of high-skilled jobs concerned with development and implementation of the new technologies, but also potentially new jobs in various sectors that are the result of economic growth due to the savings that automation brings about.

[229] Hawksworth et al., 2018; World Economic Forum, *The Future of Jobs Report 2018*. Centre for the New Economy and Society, World Economic Forum, Switzerland, 2018a, http://www3.weforum.org/docs/WEF_Future_of_Jobs_2018.pdf.
[230] Manyika et al., 2017
[231] Muro, Mark, Robert Maxim, and Jacob Whiton, "Automation and Artificial Intelligence: How Machines are Affecting People and Places," *Brookings Institution*, https://www.brookings.edu/research/automation-and-artificial-intelligence-how-machines-affect-people-and-places/, 2019.
[232] Nedelkoska & Quintini, 2018
[233] Muro et al., 2019
[234]World Economic Forum, *Towards a Reskilling Revolution. A Future of Jobs for All*, 2018b, http://www3.weforum.org/docs/WEF_FOW_Reskilling_Revolution.pdf.
[235] Muro et al., 2019

In the remainder of this section, we will discuss ethical issues in relation to the potential impact of AI and automation on the labour market, focusing on the possibility of mass unemployment, and not so much on ethical issues with new or changed jobs. We will start by identifying the value of work, and how decisions about work and employment can raise ethical issues. We will then apply these insights to the impact of AI on work and the possibility of mass unemployment.

As pointed out by the European Group on Ethics in their report on the future or work, work has both an instrumental and a non-instrumental function. Its instrumental function is to the worker and his or her family a means of existence that ensures their physical and socio-economic survival. Its non-instrumental function is to contribute to the welfare of the worker, by providing satisfaction, recognition and self-esteem.[236] Although most advanced societies provide unemployed workers with unemployment benefits that cover basic needs, it is clear that finding and retaining employment is vitally important for the well-being of workers and their families.

Although policy makers and ethicists generally do not hold there to be a right to *have* work, international human rights law holds that everyone has a right to *find* work. The universal declaration of human rights holds that people have “the right to work, to free choice of employment, to just and favourable conditions of work and to protection against unemployment.”[237] The European Charter of Fundamental Rights holds that everyone has the right to engage in work and to pursue a freely chosen or accepted occupation.[238] Both legal documents can be understood to say that people have a right to work if work is available, but not a right to availability of work. The European Social Charter of the Council of Europe, however, contains articles to the effect that states have a responsibility to ensure availability of work, as well as they can. It states, in article 1 on the right to work, that the states who have signed the Charter have a duty to achieve and maintain as high and stable a level of employment as possible.[239]

A possible outcome of AI-driven automation is that (certain classes of) low-and middle-skilled workers have a permanent low probability of finding employment due to a persistent shortage of low-and middle-skilled jobs, and them not being in a position to gain the educational and skills level to take up high-skilled jobs. The argument could be made, albeit controversially, that this situation violates their right to work, since it is an artificially created situation: there is, in fact, enough work, only employers choose to have it carried out by machines rather than human workers. Even if this argument is not accepted, the situation is clearly undesirable from a societal point of view, as the well-being and socio-economic interests of large groups in society are harmed, inequality in society is exacerbated, and social stability is undermined, and also given the agreements in the Social Charter, European states have a strong obligation to address the situation. If it turns out that certain social groups are disproportionally represented amongst those who cannot find work, then additional justice and equality arguments come into play. Moreover, past studies have shown that automation affects different regions unequally,[240] and this is also likely to happen with the current wave of automation.

[236] EGE, *Future of Work, Future of Society,* 19 December 2018. https://ec.europa.eu/info/sites/info/files/research_and_innovation/ege/ege_future-of-work_opinion_122018.pdf

[237] UN General Assembly, "Universal declaration of human rights," *UN General Assembly* 302.2, 10 December 1948, 217 A (III). https://www.ohchr.org/EN/UDHR/Documents/UDHR_Translations/eng.pdf.

[238] European Union, *Charter of Fundamental Rights of the European Union*, December 18, 2000, 2000/C 364/01. https://www.europarl.europa.eu/charter/pdf/text_en.pdf., art. 15.1

[239] Council of Europe, *European Social Charter (Revised)*, 3 May 1996, ETS 163. https://rm.coe.int/168007cf93.

[240] Martin, Ron, and Philip S. Morrison (eds.), *Geographies of labour market inequality*, Routledge, London and New York, 2003.

Several responses have been proposed to potential long-term mass unemployment resulting from AI- and robotics-driven automation. First, reskilling and retraining programs have been proposed. However, if there simply are not enough jobs to fill, and if some of the jobs will require a skill level that is unattainable for many of the unemployed, then such programs will only have limited effect.

Second, economic redistribution policies have been proposed, through taxation and subsidies. The most well-known of these is the so-called robot tax, a tax on the introduction of use of AI and robots. Such a tax could encourage employers to retain human employees by making human labour more competitive, and the gained revenue may be used to compensate those who are negatively affected by automation. However, such a tax is controversial, and arguments have been made that it is unjustified and will be ineffective.[241] Korinek (2019) argued that other ways of taxing the beneficiaries of automation may be more effective as a redistribution policy.[242] Korinek and Stiglitz (2019) suggest a combination of taxation, anti-trust policies, changes in intellectual property rights, and increased public research as redistribution measures.[243]

Third, proposals have been made to delink social protection from employment. Social protection is now strongly linked to employment, as it includes measures like unemployment insurance, income support for the unemployed, employment services, and job training, among other measures. However, if unemployment becomes a permanent condition for many, then social protections may be needed that are not related to employment, including assistance and income support not directed at employment, and the provision of universal basic services (UBS) like healthcare, education, and transportation. Most controversially, universal basic income (UBI) has been proposed. UBI involves a regular, universal and unconditional cash payment by the state, sufficient to meet basic needs, delivered to all individuals without means test or work requirement. UBI could eliminate the stigma associated with unemployment and could be a means of remunerating unpaid domestic and volunteer labour.[244] It has, however, also been criticised for disincentivising work and being too costly.[245]

Fourth, it has been proposed that cooperatives are formed, which are stakeholder – as opposed to shareholder – enterprises, that are jointly owned and governed by stakeholders.[246] Cooperatives are more likely to operate on principles of solidarity and are more likely to support workers. While there has been a recent increase in the number of cooperatives, a limitation of them as a comprehensive solution to mass unemployment due to AI automation is that if such automation gives regular private firms a competitive advantage, which seems likely, then cooperatives may not be able to compete with them if they retain their workers rather than automate.

[241] Englisch, Joachim, "Digitalisation and the Future of National Tax Systems: Taxing Robots?," *Available at SSRN:* https://ssrn.com/abstract=3244670, September 5, 2018.

[242] Korinek, Anton, "Labor in the Age of Automation and Artificial Intelligence," *Economics for Inclusive Prosperity*, February 2019. https://econfip.org/policy-brief/labor-in-the-age-of-automation-and-artificial-intelligence/.

[243] Korinek, Anton, and Joseph E. Stiglitz, "Artificial intelligence and its implications for income distribution and unemployment," in Goldfarb, Avi, Joshua Gans, and Ajay Agrawal (eds.), *The Economics of Artificial Intelligence: An Agenda*, University of Chicago Press, 2019.

[244] Giannelli, Gianna C., Lucia Mangiavacchi, and Luca Piccoli, "GDP and the value of family caretaking: how much does Europe care?," *Applied Economics,* Vol. 44, No. 16, 2012, pp. 2111-2131.

[245] Acemoglu, D. (2019). Why Universal Basic Income Is a Bad Idea. *Project Syndicate*, June 7, 2019. https://www.project-syndicate.org/commentary/why-universal-basic-income-is-a-bad-idea-by-daron-acemoglu-2019-06

[246] Borzaga, Carlo, Gianluca Salvatori, and Riccardo Bodini, "Social and Solidarity Economy and the Future of Work," *Journal of Entrepreneurship and Innovation in Emerging Economies,* Vol. 5, No. 1, 2019, pp. 37-57. https://www.ilo.org/global/topics/cooperatives/publications/WCMS_573160/lang--en/index.htm.

Fifth and finally, there is the option of growing the public sector with new types of paid public work. If the private sector is unable or unwilling to provide the number of jobs needed to avoid mass unemployment, and if mass unemployment is seen as a socially unacceptable option, then growing the public sector may be the only remaining solution. New jobs could center around the realization of public goods such as taking care of children, the elderly and vulnerable groups, environmental work, community work, and other types of jobs that contribute to society, mirroring the types of work performed in volunteer work and civil society organisations. Alternatively, and perhaps to better effect, governments can subsidise collectives and organisations initiated by people themselves to serve the public good.

Transparency and explainability

Transparency is a principle that is often demanded of artificial intelligence. It is the principle that the purpose, inputs, and operations of AI programs and algorithms should be knowable to its stakeholders so that they can understand how their decisions are arrived at. An algorithm is transparent when we understand its workings. The opposite is that it is opaque, meaning that it is a black box of which we see input/output relations but do not know how and why they are produced.

Transparency is often related to three other phenomena: interpretability, traceability and explainability. Authors relate these terms to another in different ways, sometimes distinguishing between them and sometimes equivocating them. Most importantly, *explainability* is often seen as a component of transparency. It is the ability to explain in human terms why an algorithm arrived at the decision or result. *Traceability* is the ability to use algorithmic tracing: a method for hand-simulating the execution of a program-coded algorithm in order to manually verify that it works correctly before it is compiled. *Interpretability* is given different meanings, some of them identical to transparency, some to explainability, and some different from both (see also the section on machine learning).

Preece et al. (2018) have argued that the diversity in definitions of interpretability, transparency and explainability is due to an inability to distinguish the different stakeholder communities in relation to which they are defined.[247] They distinguish four stakeholder communities: developers (people who build AI applications), theoreticians (people concerned with AI theory, particularly around neural networks), ethicists and users. They argue that these stakeholders have different capabilities and different needs for transparency and explainability, resulting in different conceptions. They do not argue, however, that different types of explanations necessarily have to be developed to satisfy them, arguing that it is also possible to develop composite explanations that contain information for multiple stakeholders and that can be unpacked per a stakeholder's particular requirements.

Lack in transparency is especially an issue in machine learning algorithms. Lipton, focusing on such algorithms, distinguishes between transparency, which he defines as grasping how a model works, and post-hoc interpretability, which is the explanation of an algorithm's output without appeal to its inner mechanics, often through verbal explanations and visual aids.[248] For example, it explains why a neural network classifies an object as a tumour by referring to its similarity to other objects it has classified as tumours. He also argues that lack of transparency may be caused by the complexity of the algorithm, which can happen in any of three ways: the output cannot be replicated by a human, some features of the algorithm are too complex, or the *type* of algorithm is simply uninterpretable. Especially the latter

[247] Preece, Alun, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty, "Stakeholders in explainable AI," *arXiv preprint arXiv:1810.00184*, 2018s.

[248] Lipton, Zachary C., "The Mythos of Model Interpretability," *Communications of the ACM,* Vol. 61, No. 10, 2018, pp. 36–43.

has caused reason for concern. Some types of algorithms, such as neural networks, especially deep learning algorithms, are to a large extent black boxes. These algorithms provide no way of explaining how they reach their results, contrary to for instance linear algorithms that always "converge to a single solution."[249] Neural networks are therefore criticized for their inability to allow for explanations of their computation due to their hidden layers.[250] Furthermore, unsupervised learning algorithms are even less transparent due to a lack in data labels. This makes it harder to analyse these algorithms.[251]

Besides technical challenges that complicate the possibility of full transparency, explanations may be too technical to be understood by a lay community. One may argue that full transparency leads to the "'one true' explanation" as post-hoc explanations are not intrinsic in an algorithm; they tend to be only possible with specific training.[252] Nonetheless, due to the black box characteristics of deep learning algorithms, explainable AI (XAI) is focused on such post-hoc explanations.[253] An example of a post-hoc explanation uses similar examples (e.g., pictures) to show why the algorithm came to its conclusion. DARPA has launched a project to improve explainability of AI based on post-hoc explainability.[254] Generally, a system that is explainable should support several resulting traits, which are confidence, trust, safety, ethics and fairness.[255] XAI is beneficial as it "may allow more efficient and effective use of the technology",[256] and a system that can explain itself may be considered more trustworthy by outsiders.[257] Miller, Howe and Sonenberg warn for the possibility that XAI will still only be explainable for the developers, rather than for the users.[258] Therefore, they advocate for an integration of algorithmic models with social sciences.

The European Commission's High-Level Expert Group on Artificial Intelligence's ethics guidelines advocate a principle of transparency for AI systems, and divide it up into three components: traceability, explainability and communication.[259] As noted earlier, *traceability* means that the data sets and the processes and algorithms that yield an AI's decision, including the processes of data gathering and data labelling, should be documented well to allow for traceability and an increase in transparency and explainability. *Explainability* includes *technical explainability*, which is the type of explainability referred to earlier and relates to the ability to understand why a system reached its decision, and *business model explainability*, which is the availability of accounts of the purpose and function of the system within an organisation, and its influence on organisational (decision-making) processes. *Communication*, finally, means that users are properly informed about the fact that they

[249] Ibid., p. 40.
[250] Hildebrandt, Mireille, and Serge Gutwirth, "Concise conclusions: Citizens out of control," *Profiling the European Citizen*, Springer, Dordrecht, 2008, pp. 365-368.
[251] Garg, Vikas K., and Adam Tauman Kalai, "Meta-Unsupervised-Learning: A supervised approach to unsupervised learning," *arXiv preprint arXiv:1612.09030*, 2016.
[252] Sheh, Raymond, and Isaac Monteath, "Defining Explainable AI for Requirements Analysis," *KI-Künstliche Intelligenz,* Vol. 32, No. 4, 2018, pp. 261-266., p. 263
[253] Ibid.
[254] Gunning, David, "Explainable artificial intelligence (xai)," *Defense Advanced Research Projects Agency (DARPA), nd Web,* Vol. 2, 2017.
[255] Doran, Derek, Sarah Schulz, and Tarek R. Besold, "What does explainable AI really mean? A new conceptualization of perspectives," *arXiv preprint arXiv:1710.00794*, 2017.
[256] Brinton, Chris, "A framework for explanation of machine learning decisions," In *IJCAI-17 Workshop on Explainable AI (XAI)*, 2017, pp. 14-18., p. 14
[257] Fox, Maria, Derek Long, and Daniele Magazzeni, "Explainable planning," *arXiv preprint arXiv:1709.10256*, 2017.
[258] Miller, Tim, Piers Howe, and Liz Sonenberg, "Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences," *arXiv preprint arXiv:1712.00547*, 2017.
[259] High-Level Expert Group on Artificial Intelligence (2019). Ethics Guidelines for Trustworthy AI.

interact with an AI system rather than a human being, and that they are informed about the system's capabilities, limitations, and level of accuracy.

Let us now turn to an ethical discussion of transparency. Three arguments have been presented for the moral importance of transparency in AI (and of explainability and other related concepts). The first is that transparency is needed to protect the rights and interests of those who are affected by the system. The decisions resulting from AI systems can have serious consequences for people's interests and rights. People, it is claimed, have a right to know why the system makes the decisions it does, so that they can assess the fairness and reasonableness of these decisions, and challenge the reasons for these decisions if they do not seem sound.[260] This is especially important for those decisions that affect people's fundamental rights or that significantly affect their interests. The High-Level Expert Group on Artificial Intelligence stated: "Whenever an AI system has a significant impact on people's lives, it should be possible to demand a suitable explanation of the AI system's decision-making process."[261]

A second argument is that transparency, and related notions, are needed in order to ensure (algorithmic) accountability for AI systems.[262,263] (More on the ethical issues in relation to algorithms and responsibility in subsection 5.1.2 and subsection 5.1.3, respectively.) The argument here is that it should be possible for organisations to assume responsibility for the AI systems they use, and that it should be possible for others to hold such organisations accountable. But responsibility and accountability are severely restricted if both the organisation and its stakeholders are not able to determine why its systems make the decisions they make, and to correct and improve decision-making procedures that are considered erroneous or unfair. If one is to have algorithmic accountability, it seems, transparency seems to be a necessary condition that ought to be in place.

Third, it has been claimed that transparency is needed in order to ensure trust in AI by its users and stakeholders. This is claimed by the HLEG, which holds that it "is crucial for building and maintaining users' trust in AI systems" (p. 13).[264] Trustworthy AI, for the HLEG, is AI that is reliable and functions sufficiently in the interest of users and affected stakeholders. If users and other stakeholders cannot verify that the technology functions correctly and to their benefit, they will be less inclined to trust it, and this may lead to resistance, avoidance, and improper usage.

While all three arguments appear to have a degree of validity, arguments have also been developed that too strong a demand for transparency could lead to unnecessary costs and could limit innovation. Transparency, after all, comes at a cost. It will require significant investments to develop new concepts of transparency and implement them in AI systems, a demand for transparency may limit other aspects of the performance of the systems such as accuracy because trade-offs are made in design, the demand for transparency may conflict with intellectual property rights of developers of AI systems, and a strong requirement of transparency may mean, at least for now and perhaps indefinitely, applications of neural networks and machine learning may have to be limited, as there are no good approaches for ensuring their transparency.

[260] Ibid.

[261] Ibid., p. 18.

[262] Ananny, Mike, and Kate Crawford, "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability," *new media & society,* Vol. 20, No. 3, 2018, pp. 973-989.

[263] Diakopoulos, Nicholas, "Accountability in algorithmic decision making," *Communications of the ACM,* Vol. 59, No. 2, 2016, pp. 56-62.

[264] High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*, European Commission, 2019. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419. To be precise, the HLEG makes this claim in relation to the principle of explicability, which it claims to be closely related to transparency. It does not, however, explain what explicability is and how it is different from transparency.

Zerilli et al. (2018) argue that proponents of transparency may be setting automated decision-making to an unrealistically high standard, a standard to which human decision-makers could not be held.[265] For human decisions, we accept "intentional stance" explanations that appeal to mental states, and we do not require mechanistic explanations that lay bare the underlying causal mechanism. He argues that there are few circumstances in which it is justified to hold AI to a higher standard than humans.

In the trade-off between transparency and cost, also, it appears that transparency is more important for some applications than for others. Clearly, transparency is not very important for an AI program that controls the movements of a pet robot. Clearly, it is important for a program that recommends sentencing guidelines for felonies. Then people have a strong interest in knowing why a program made the recommendations it did, so that they may either accept or appeal them. A differentiated approach seems justified that allows for different degrees and kinds of transparency for AI systems relative to the interest of stakeholders and society in understanding, evaluating, appealing and correcting the decisions of AI systems and holding its organisational developers and users accountable.

Other potential harms

Besides the issues that have been described thus far, there may be other potential harms as a result of AI technology that deserve to be mentioned. These include, amongst others, potential harms to the meaningfulness of work and life, harms to democracy, and harms due to a misplaced sense of trust in AI systems. Let us briefly discuss these three issues.

First, AI technology may have a negative impact on the meaningfulness of work and life. Generally speaking, people do not work simply to earn a salary; they also work because it adds meaning to their lives and gives them a sense of purpose.[266] Meaningful work has been defined as "that which actualises human potentials [including] creativity, autonomy, abilities and talents, identity, sociality, and is necessary to fulfil a human end or purpose, e.g., happiness, self-development and well-being, or personal development."[267] Our work should be "structured by the core goods of freedom, autonomy and dignity," in order to achieve a "sense of being a vivid presence in collective action."[268] Through the introduction of AI technology, we may run the risk that many jobs are becoming increasingly mundane, circumscribed and controlled as AI technology removes the need for specific human skills that had previously been required for those jobs.

In addition, people may struggle to find meaning when they lose their jobs permanently as a result of AI technology. Even if AI technology offers people opportunities to have a sense purpose outside of work, it is not clear whether they will actually be ready for or be able to conceive of a meaningful life that is completely disconnected from work, and in particular the kind of job they had devoted a large part of their life to.[269] Moreover, finding meaning in daily life outside of work may also become increasingly hard as AI technology is not only likely to take over jobs, but also a range of personal and domestic roles. These roles may include parenting, elderly care, volunteer work, as well as other forms

[265] Zerilli, John, Alistair Knott, James Maclaurin, and Colin Gavaghan, "Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?," *Philosophy & Technology*, 2018, pp. 1-23.

[266] Beard, Matthew, "With robots, is a life without work one we'd want to live?," *The* Guardian, September 26, 2016. https://www.theguardian.com/sustainable-business/2016/sep/26/with-robots-is-a-life-without-work-one-wed-want-to-live

[267] Harnden, Charlie, "How Artificial Intelligence is Destroying Meaningful Work," *Medium*, https://medium.com/@charlieharnden/artificial-intelligence-and-meaningful-work-c8f6ec24f11b

[268] Yeoman, Ruth, "Can Artificial Intelligence Give Our Lives Meaning?," AIA News, Issue 69, June 26, 2018. https://iai.tv/articles/can-ai-generate-meaning-in-our-lives-auid-1101

[269] Beard, 2016, op. cit.

of support for vulnerable people.[270] Finally, it has been argued that if we outsource the most intellectually demanding, frustrating and boring jobs to AI systems, this might diminish our appetite for stimulating and meaningful work and reduce our skills to spend our time meaningfully.[271]

Second, AI technology may pose a threat to democracy. A host of AI techniques may be deployed by political and non-political actors in ways that can ultimately undermine democracy, notably including (1) filtering and/or restricting people's access to information, (2) manipulating voters through the production and dissemination of misinformation, and (3) suppressing people's freedom of speech. Through the use of sophisticated filtering algorithms, social media platforms and search engines currently have a major role in determining what and how politically relevant information is taken in by people. Recently, social networks (e.g., Facebook, Twitter) have been accused of using algorithms that lead to the creation of filter bubbles, and may favour populism as they prioritize content that engages users the most.[272] Arguably, a healthy democracy requires that its citizens are able to consider a wide variety of information and viewpoints, the selection and presentation of which are free from bias.

Furthermore, AI technology may also make it easier to manipulate voters through AI-based creation and dissemination of propaganda and misinformation. AI techniques have been deployed on social media to "micro-target" and profile voters on the basis of their personal data, using personalised messaging and fake news to compel them to vote a certain way.[273] In addition, AI techniques have been used to create so-called *deepfakes*, where important parts of videos and images such as people's faces are replaced with (parts of) other videos or images without a resultant loss of apparent realism. Such deepfakes have recently started to be used for political purposes.[274] (See also the previous discussion on "Dual use and misuse" in this subsection.) Also, AI techniques allow for the creation of virtual agents or "bots" that can flood the comment sections of websites and social media with complementary or disparaging comments and popularize content from extremist, sensationalist and conspiratorial sources, thus influencing public opinion.[275]

Moreover, AI technology also forms the basis of sophisticated methods to directly suppress freedom of speech and restrict access to information, such as through Internet censorship. In China, for example, AI techniques are being used to guard the borders of its "Great Firewall" and to stifle free speech by shutting down material that the Chinese government deems objectionable.[276] What may further compound these issues in relation to democracy is that significant and largely unchecked powers to utilise and control AI technology in social spaces tend to concentrate in few very large technology companies (e.g., Google, Amazon, Facebook, Apple, IBM) and state-owned enterprises (e.g., in China), which possess vast resources to develop and improve their data-hungry AI systems.

Third and finally, AI technology may pose ethical issues in terms of trust. AI systems' superiority to humans at performing certain tasks can instil in people a false belief in the infallibility of these systems.

[270] Ibid.

[271] Ibid.

[272] Bernard, Pascal, "Is AI a threat to Democracy?," *Towards data Science*, May 21, 2019. https://towardsdatascience.com/is-ai-a-threat-to-democracy-4bef3e5fcfdd

[273] Ghosh, Dipayan, "What is microtargeting and what is it doing in our politics?," *Internet Citizen*, October 4, 2018. https://blog.mozilla.org/internetcitizen/2018/10/04/microtargeting-dipayan-ghosh/

[274] Romano, Aja, "Jordan Peele's simulated Obama PSA is a double-edged warning against fake news," *Vox*, April 18, 2018. https://www.vox.com/2018/4/18/17252410/jordan-peele-obama-deepfake-buzzfeed

[275] Howard, Philip, "How Political Campaigns Weaponize Social Media Bots ," *IEEE Spectrum*, October 18, 2018. https://spectrum.ieee.org/computing/software/how-political-campaigns-weaponize-social-media-bots

[276] Murison, Malek, "The Great Firewall: China looks to AI to censor online material," *Internet of Business*, May 23, 2018. https://internetofbusiness.com/china-censorship-online-material-ai/

Even the most sophisticated AI systems cannot be trusted to achieve 100 percent perfect results all of the time. Overconfidence in the results of these systems can amplify some of the potential harms that have previously been described, including issues of safety and justice. Considering the breadth of AI techniques currently in existence, their often complicated and opaque nature, and the fact that many have not been sufficiently tested, it is still unclear how much should we can generally trust AI systems.[277] Misplaced trust may have disastrous consequences, especially in cases where statistics-based learning systems are used, which are known to occasionally produce bizarre outlier mistakes. Since humans would hardy make such mistakes, it can be hard to predict them and adequately guard against them.

5.2. General ethical issues in robotics

This subsection offers a discussion of the general ethical issues in robotics. We begin, in subsection 5.2.1, by describing the ethical issues that are inherent in the general aims of robotics and its subfields. Then, in subsection 5.2.2, we detail for the most important robotics techniques, methods and approaches, the main ethical issues that are specific to them (i.e., issues that are inherent in, or frequently occur with, these techniques, methods and approaches). Finally, in subsection 5.2.3, we describe the main ethical issues with regard to some of the general implications and risks of robotics technology (e.g., harms to autonomy, privacy, justice). Figure 3 provides an overview of the structure of these three subsections.

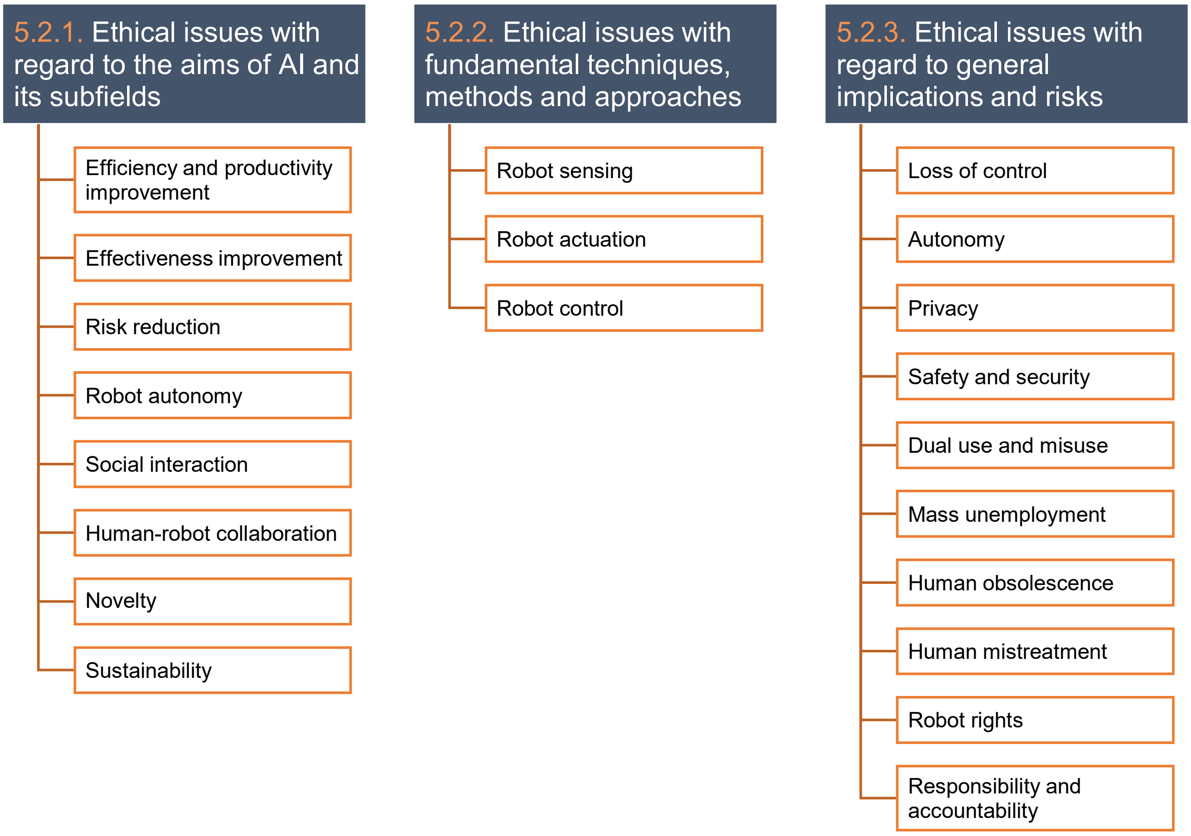

5.2.1. Ethical issues with regard to the aims of AI and its subfields
Efficiency and productivity improvement
Effectiveness improvement
Risk reduction
Robot autonomy
Social interaction
Human-robot collaboration
Novelty
Sustainability
5.2.2. Ethical issues with fundamental techniques, methods and approaches
Robot sensing
Robot actuation
Robot control
5.2.3. Ethical issues with regard to general implications and risks
Loss of control
Autonomy
Privacy
Safety and security
Dual use and misuse
Mass unemployment
Human obsolescence
Human mistreatment
Robot rights
Responsibility and accountability

Figure 4: Structure of subsection 5.2 on general ethical issues in robotics.

[277] Hurlburt, George, "How Much Should We Trust Artificial Intelligence," *InfoQ*, September 8, 2017. https://www.infoq.com/articles/ai-trust/

5.2.1. Ethical issues with regard to the aims of robotics and its subfields

In this subsection, we identify and analyse the ethical issues associated with the most important aims and sub-aims in the development of robotics systems. We have identified the following ethically relevant aims and sub-aims of robotics: *efficiency and productivity improvement*; *effectiveness improvement*; *risk reduction*; *robot autonomy*; *human-robot collaboration*; novelty; and *sustainability*. For each of these, we discuss below the most important ethical issues.

Efficiency and productivity improvement

One of the main goals or desirable outcomes of utilizing robots is that of efficiency.[278,279,280] The general idea of efficiency goals in robotics surround achieving more, or better, results for less time, money, effort, and/or risk. Different fields of robotics have relatively different conceptions of what efficiency means and looks like to them, but they can seemingly be classified into two dominant forks: efficiency (optimizing, downsizing, capital),[281] and effectiveness (performance, enhancement, gap-filling).[282] *Optimizing and downsizing* efficiency goals are classified as such when the main aim is achieving greater efficiency by reducing the number of human labourers or interactors, tools, techniques, or space. To exemplify: if a manufacturing plant requests a robot that is able to not only cut parts from metal sheets, but also measure, grind, and solder certain ones together, the factory's goal is to become more efficient by reducing labourers needed (as the robot can do four tasks by itself), training cost/time (cutting, grinding, measuring, and welding are all distinct tasks with few individuals able to do all, especially simultaneously), space (instead of separate areas and benches, all tasks can be done in one area), and tools (the robot can measure and cut without additional tools required). *Performance and enhancement* efficiency goals can be classified as such when the main aim is achieving greater efficiency by enhancing the range of human, tool, or technique performance or performing tasks humans could not, should not, or are difficult for them to achieve. On such example of this is exoskeleton use in manufacturing another is UAVs trying to pinpoint the source of the Fukushima nuclear radiation. This fork of efficiency does not mind so much if there is space, time, or quantity of people reductions, but rather focuses on improvement of task completion, success, and performance measures.

The immediately apparent benefits of efficiency goals in robotics is exactly what has been stated: money saved, faster task completion, broader task completion, and risk reduction. On one hand, this could be an excellent opportunity to prevent human labourers from toiling endlessly on repetitive, mundane, or dangerous tasks and open-up different labour opportunities surrounding machine collaboration, maintenance, and growth. On the other hand, especially in the case of optimizing/downsizing efficiency goals, potential areas of concern are mass layoffs and overspecialization of entry-level labour (less need for manual or uneducated labour, greater need for

[278] Rama Fiorini, Sandro, Bermejo-Alonso, et al., "A Suite of Ontologies for Robotics and Automation", *IEEE Robotics & Automation Magazine*, March 2017.
[279] Bhoge, Anand, "Smart Robotics: Revolution is Motto, Efficiency is Aim", *Robotics Tomorrow*, August 2018.
[280] Denicolai, Lorzenzo, Grimaldi & Palmieri, Silvia, "Videos, Educational Robotics and Puppets: An Experimental Integration of Languages", Universita Degli Studi Di Torino, 2017.
[281] Focused Ultrasound Therapy Using Robotic Approaches (FUTURA) Project, "Project Objectives", accessed December 2018.
[282] Nezhadali, Vaheed, "Multi-Objective Optimization of Industrial Robots", Linköpings Universitet, 2011.

engineers, scientists, etc.).[283,284,285] Of further concern is increasing expectations for workers that are already there, but no increase of training or pay. This leaves general labourers trying to "make it work" with robotic co-workers despite not having the knowledge on how to maintain, repair, or calibrate them. Thus, robots in some of these contexts end up underutilized and underperforming on efficiency benchmarks.

Cost reduction/profit increases are also highly desirable outcomes of implementing robots.[286,287,288,289] While the initial cost of integrating robots into a workspace might be high, the return on investment (ROI) is seemingly well worth the initial splurge. Not only are robots able to work constantly without breaks, health insurance, or long-term injuries, they are able to reduce labour costs significantly—some machines only costing €1.70-€2.60 per hour including maintenance costs. This is drastically able to increase profit margins for any company desiring to make use of robots while improving workflow, increasing floor space, and reducing costly mistakes and risks.

Unfortunately, if companies obsess too much over profit margins and increase demand for faster, more efficient, and independent machines, risks and dangers to human workers increase due to lowering of machine awareness. Fully automated workplaces, known as "lights out" workplaces, have yet to be fully adopted from fledgling stages and remain frequently debated if they should be an approach to labour humans ought to be striving for.[290] As such, machines will have to have a speed/efficiency cap to allow easier interfacing for humans and other less advanced equipment.

Effectiveness improvement

One of the other main goals of robotics that closely ties to efficiency is that of compensation. Compensation, or "gap-filling" refers to goals seek to fulfil or assuage shortcomings in humans, resources, or environmental capabilities.[291] Examples of such a goal can be seen in healthcare, with caretaking robots easing the burden of high healthcare demand and inability to satisfy the demand due to cost, coverage, or caretakers. Robots can be implemented to take care of some of the routine activities to help caretakers focus on the more pressing challenges (delivering clean materials, giving

[283] Manyika, James, Lund, Susa & Chui, Michael et al., "Jobs Lost, Jobs Gained: What the Future of Work Will Mean for Jobs, Skills, and Wages", McKinsey & Company, November 2017.

[284] Decker, Michael, Fischer, Martin & Ott, Ingrid, "Service Robotics and Human Labor: A First Technology Assessment of Substitution and Cooperation", *Elselvier Robotics and Autonomous Systems* 87, January 2017.

[285] Owais Quereshi, Mohammed & Sajjad Syed, Rumaiya, "The Impact of Robotics on Employment and Motivation of Employees in the Service Sector, with Special Reference to Health Care", *Oshri Saftey and Health at Work* 5(4), December 2014

[286] Brown, Robert, "Robots Make A Money-Making Assembly Line by Cutting Costs", *Center for The Future of Work*, February 2015.

[287] RobotWorx, "How Can Industrial Robots Improve My Profits?" Accessed December 2018. robots.com/faq/how-can-industrial-robots-improve-my-profits

[288] Carlisle, Brian, "Pick and Place for Profit: Using Robot Labor to Save Money", *Robotics Business Review*, September 2017.

[289] The Boston Consulting Group, "The Shifting Economics of Global Manufacturing", *February 2015*.

[290] Shedletsky, Anna-Katrina, "When Factories Have a Choice, It's Best to Start with People", *Forbes*, June 2018.

[291] Vega, Julio & Canas, Jose, "PiBot: An Open Low-Cost Robotic Platform with Camera for STEM Education", *Electronics,* December 2018.

directions, moving patients)[292]. Other examples of this can be seen in UAVs helping to patrol coral[293], drones planting seeds[294], or robots helping to manage crops.[295]

The negative aspects of this goal are not well documented. Unlike efficiency, the argument against labour replacement is not as strong—as robots, in this case, would be seeking to help humans compensate in areas of need. Especially with resource and environmental constraints, these tasks simply would be left undone if robots were not constructed to do them, leaving the lives of humans and other environmental inhabitants seemingly poorer in the process. Two potential avenues for criticisms could be (1) an overreliance upon robotic assistants. Humans may become so dependent upon "technofixes" for environmental/resource constraints that these fixes reduce our willingness to address the root causes of these problems (e.g., overconsumption, poor resource management, industrialization). This could potentially lead to a slippery slope of worsening conditions until even technology is not enough to resolve the issues humans have caused. Further, outsourcing too many tasks may lead to increased job performance expectations that could decrease labourer satisfaction and morale. Additionally, (2) increasing the quantity of robots to fill gaps without effective and efficient recycling and materials could potentially cause more problems than they resolve.

Risk reduction

Risk management and safety assurance are other commonly referenced goal in robotics.[296,297] How this is achieved, precisely, seems to vary widely between robotics fields. For example, risk management for healthcare robotics may revolve more tightly around security, accuracy, privacy, and patient outcomes; risk management in manufacturing might encompass values such as precision, speed, situational awareness, communication, collision avoidance, and failsafe mechanisms. Which values the field prizes as goals highly depends on what is seen as a "risk" in that field already, what the robot is tasked with accomplishing, and how it is designed. The risk management for a Roomba will be significantly less complicated than that of Softbank's NAO, the healthcare assistant.

This goal is seemingly a noble one. It does not spark much controversy either in robotics disciplines or related ethics fields, as a general goal. The main points of contention of this goal come to light in debates around *how* risks need to be assessed surrounding the implementation of this technology. Various risk assessment methodologies exist, various fields present different risks and assessment demands, to be sure, but the overarching goal of preventing harm or potential harm seems to be a good place to start. The other challenges risk management goals face is that of longevity and horizon scanning—ensuring prevention of the "hydra effect"[298] for long term implementations and a more holistic account of risk management that considers various angles of risk that may not be readily apparent. A few to think about could be mass job loss, decreased worker morale, civil unrest, environmental harm, et cetera.

[292] Aetheon. "Tug Informational Graphics", 2018. Accessed December 2018. aethon.com/infographics
[293] Braun, Ashley. "The RangerBot is a New Line of Defense Against Coral-Eating Crown-of-Thorns Starfish", Smithsonian, August 2018.
[294] Droneseed, "Precision Forestry", accessed December 2018. droneseed.co
[295] Wall-ye, "MYCE_Vigne", accessed December 2018. wall-ye.com/index-2.html
[296] Stephens, Tim, "Robotics Project Aims to Develop Systems for Human-Robot Collaboration", *UC Santa Cruz Newscenter*, December 2012.
[297] P. Lichocki, P. Kahn Jr, and A. Billard. The Ethical Landscape of Robotics. *IEEE Robotics and Automation Magazine*, 18(1):39-50, 2011. (Cited as requested)
[298] The "hydra effect" refers to the counter-intuitive situation when actions to reduce a particular problem actually stimulate its multiplication.

Robot autonomy

While the goal of robots operating without human intervention has lost some of its popularity in robotics fields like manufacturing and healthcare, autonomy is still a common sub-goal when it comes to optimization, compensation, and risk reduction.[299,300] Other areas in which self-sufficient robots are desirable are in tedious, repetitive, or monotonous tasks, vacuuming, sorting, or delivering objects, to exemplify.[301] Autonomy is also a desirable goal when it comes to the topics mentioned in the Compensation category, as robots that can emulate, to a reasonable degree, in certain social contexts, such as education, home care, and social relations, may be able to fill critical roles humanity is unable to accommodate.[302,303]

The ambiguity of long-term effects more autonomous robots create is cause for concern across multiple robotics fields. Some such concerns are rooted in the social, with respect to the loss of interpersonal interactions and interdependence between human beings. Further concerns highlight the loss of human dignity or a decrease in quality of service human labourers provide.[304] Alongside quality concerns is that of safety and control, more autonomous machines potentially leading to human labour obsolescence and deskilling as well as difficulty holding parties accountable when accidence to happen. Although results have been to the contrary—while job loss with technological growth has been seen, labour has managed to adapt and rebalance with needs being created elsewhere at each bump up of advancement.[305] Especially in social robotics, it is important to assess whether the use of robots is better than nothing at all, even facing potential unknowns. Further, some proponents of AI and robotics argue that some of these 'unknowns' of dedicating tasks to autonomous robots might very well be beneficial in ushering humanity into a new age.[306]

Human-robot collaboration

As a more recent development in response to both the compensatory need for robotic labour and the awareness of the potentially perilous socio-technical pitfalls of automation goals, collaborative robots, AKA 'cobots', seem to provide a halfway-happy transition point for each side of the debate. Collaboration goals in robotics seek to design and integrate robots that are not merely able to function alongside humans, but to actively participate in task completion with humans. It is seemingly the hope of these goals in robotics not to design the fastest, smartest, most autonomous, money-generating

[299] Peertechz Journal, Engineering Group, "Aims and Scope", *Annals of Robotics and Automation*, accessed December 2018.
[300] Mainbot, "Industrial Objectives", accessed December 2018.
[301] Bekey, George A., *Autonomous Robots: From Biological Inspiration to Implementation and Control*, MIT Press, February 2017.
[302] Pachidis, Theodore, Vrochidou, Eleni, & Kaburlasos, Vassilis et al., "Social Robotics in Education: Stat-of-the-Art and Directions, 27th International Conference on Robotics RAAD, July 2018
[303] Foster, Malcolm, "Aging Japan: Robots May Have Role in Future of Elder Care", *Reuters,* March 2018.
[304] Dhir, Amandeep, Yossatorn, Yossiri, Kaur, Puneet, & Chen, Sufen, "Online Social Media Fatigue and Psychological Wellbeing- A Study of Comulsive Use, Fear of Mission Out, Fatigue, Anxiety and Depression", *Elselvier International Journal of Information Management*, June 2018.
[305] Vardi, Moshe, "What the Industrial Revolution Really Tells Us About the Future of Automation and Work", *The Conversation*, September 2017.
[306] Mortensen, Dennis, "Automation May Take Our Jobs—But It'll Restore Our Humanity", *Quartz Automation Revolution*, August 2017.

robot around, but rather to aim for job loss prevention, creative approaches to labour in the age of robotics, and finding a coexisting harmony between humans and robots.[307,308]

It seems that collaborative robots, being a solution to several problems, do not generate as many negative attitudes and reactions as other goals do. While there definitely seems to be questions circling the research community of best design practices, contexts, and costs, these questions and concerns stay relatively well in the realm of the pragmatic for many of the more moderate-leaning robotics supporters. Those adamantly opposed to heavy use of technology for compensating for human inability will probably still call "technofix" fouls for this solution, and those heavily in favour of increasing profit margins and increasing productivity will still find full-automatic a more appealing goal. However, collaborative robotics, when sought after with candour, awareness, and creativity, may yield more beneficial results than could be anticipated for humans and robots alike.

Novelty

Novelty goals in robotics are being categorized as aims to entertain, innovate, or create for their own sakes. These types of robots may not have any further purpose than to see if they can be built, to bring joy, or to garner attention and act as art. Examples of these goals being executed in the wild are the GogBot festival in Enschede, The Netherlands,[309] the designs of robotic artist Jan de Coster at Slightly Overdone Studio,[310] or WowWee's MiP robotic toy.[311]

While these types of goals seem quite innocent and low risk for negative outcomes and/or impact, one of the lesser noted consequences of these types of robotic goals is that of e-waste. With quickly paced upgrades and the notion of having the "latest and greatest" technologies, these types of toys seemingly fall into obsolescence from year to year, and none of the companies creating them share a waste management or recycling program that they have in place for buying back old toys or what is done when their products do not sell.

Sustainability

Bridging onto the obsolescence of novelty goals in robotics are the more futuristic goals of sustainable robotics. These goals are not focused on the robots helping to make humans more sustainable, but rather on making the robots themselves sustainable—with biodegradable materials that heavily reduce or eliminate robotic e-waste.**[312]** These types of waste consciousness goals for the robot itself seem to be, as of 2018, entering the dialogue in robotics design. Hopefully, in the future years, more discussions about robotics recycling, outlawing planned obsolescence, and material consciousness will help to achieve environmental goals in robot design as well.

[307] Cangelosi, Angelo & Schlesinger, Matthew, "From Babies to Robots: The Contribution of Developmental Robotics to Developmental Psychology", *Child Development Perspectives,* February 2018.

[308] Zaleski, Andrew, "Man and Machine: The New Collaborative Workplace of the Future", *CNBC Tech*, October 2016.

[309] Gogbot, "About". Accessed December 2018. 2018.gogbot.nl/about/

[310] Accessed December 2018: sulu.be/SlightlyOverdone/

[311] Accessed December 2018: wowwee.com/mip

[312] A, Akhil, "Researchers Aim to Build Eco-Friendly Robots with Biodegradable Materials", *Sastra Robotics*, July 2018.

5.2.2. Ethical issues with regard to fundamental techniques, methods and approaches

In this subsection, we describe the most important fundamental techniques, methods and approaches in robotics and the main ethical issues that are specific to them (i.e., issues that are inherent in, or frequently occur with, these techniques, methods and approaches). A listing of many important techniques, methods and approaches in robotics was made in one of our previous deliverables, SIENNA Deliverable 4.1. This section provides further elaboration on some of those concepts and specifically identifies ethical concerns which may arise in relation to them. Whilst ethical concerns often only make themselves apparent at the application stage, it is still possible to make some generalisations about the fundamental technique, method or approach in question. To that end, the main areas discussed here will be sensing, actuation, and control. Sensors are needed so that the robot can obtain information from its environment; actuators are there to give the robot the ability to move and exert forces on its environment; and an on-board computational capacity is required for the robot to have some level of autonomy.

Please note that some significant approaches and subfields of robotics (such as humanoid robotics, social robotics and biohybrid robotics) are discussed in section 6.2 of this report, since for the most part the ethical issues of concern here manifest themselves in specific products. Also note that AI-related techniques, methods and approaches that are applied in robotics are discussed in subsection 5.1.2. Finally, it should be emphasised that our listing of specific techniques, methods and approaches in robot sensing, actuation and control is not exhaustive; we have only attempted to identify and discuss the most important techniques, methods and approaches that may also give rise to significant and specific ethical issues.

Robot sensing

Sensors are devices, modules, or subsystems which constitute a robots' window on to its environment. A robot might use a sensor for a range of purposes: to identify a target, to detect an obstacle, to build maps or to determine its own location. All these situations require that information be sent from the sensor to other electronic components within the robot for processing. Each sensor is based on a transduction principle, where energy is converted from one form to another – from analogue to digital and back to analogue. Sensors are vital for the robot to be able to deal with uncertainties and be an active participant in their environment.[313]

Sensors can be classified as either *proprioceptive* or *exteroceptive*. Proprioceptive sensors measure the internal state of the robots' system—battery level, wheel position or a joint angle, for example; exteroceptive sensors measures the external environment and objects in it. Sensors are also classified as either passive or active. Passive sensors, such as a camera or microphone, only receive energy from the environment; active sensors, such as radar, emit some kind of energy. In each case, the information processed is used to calculate an appropriate response and/or relay it to human operators or supervisors. Finally, sensors can also be classified as two general types: contacting and non-contacting. Contact or tactile sensors rely on things like touch and force sensing, proximity or displacement, or slip sensing.[314] Non-contacting sensors include visual and optical sensors, magnetic and inductive sensors, capacitive sensors, resistive sensing, ultrasound, sonar and air pressure.[315]

[313] Aparna, Kale. Bodhale, Umesh "Overview Of Sensors For Robotics", International Journal of Engineering Research and Technology (IJERT) Volume 02, Issue 03 (March 2013).
[314] Ibid.
[315] Ibid.

Robots can be fitted with a wide variety of sensors. Some common robot sensors include:

- *Cameras.* A camera is a device which focuses light on a photosensitive surface and captures it as a still or moving image.
- *Microphones*. Microphones convert soundwaves from the environment into an electrical signal which may then be amplified, transmitted or recorded. They may be used by a robot for navigation or for a range of other purposes.
- *Accelerometers.* Accelerometers measure the gravitational acceleration, tilt and vibration of the device they are mounted on. Inside the sensor, a MEMS device (Micro-Electro-Mechanical-System) made of tiny micro-structures bends due to momentum and gravity.
- *Thermometers*. Thermometers measure the temperature of solids, liquids or gases. They are composed of a temperature sensor and a medium which converts physical change into a numerical reading. Robots may use thermometers to monitor their internal temperature or that of their environment.
- *Vibration sensors*. Vibration sensors measure linear velocity, displacement and proximity, and acceleration. They can be a useful tool for gauging the condition of a robot.
- *Infrared sensors*. Infrared sensors measure the characteristics of an environment by emitting or detecting infrared radiation. They can measure the heat emitted by an object and detect motion. A robot may use this kind of sensor for certain tasks such as object detection and obstacle avoidance.
- *Radar*. Radar (Radio detection and ranging) sends out beam pulses of high frequency electromagnetic fields and detects reflections of the beam from nearby objects. The time taken for the signal to be sent and returned is used to calculate distance. Lidar is a type of radar commonly used in robotics which uses light pulses to detect the distance of objects. Both types of radar may be used by robots to navigate their environments.
- *Sonar*. Sonar acts on a similar principle to radar. Sonar emits a mostly inaudible sound and detects the returning echo. As with other radar systems, sonar can be used by a robot to navigate its environment and detect objects.

The use of some of these types of sensors may give rise to specific ethical issues, the first of which is privacy. Several of the sensors listed are classified as exteroceptive and passive. With sensors of this type, we can identify privacy as a potential concern since these sensors require that data be gathered from the external environment. Photographing or filming the external environment with a camera, for example, may infringe on the bodily privacy of others, especially if images are being captured indiscriminately. Similarly, microphones raise the possibility of audio data being indiscriminately collected from the environment. Private conversations amongst individuals, for example, may be captured by a microphone.

The other types of senses listed – accelerometers, thermometers and vibration sensors – do not appear to present an immediate threat to privacy. However, they could pose such a threat if combined with other data. Accelerometers, for example, provide data on movement. Depending on the robot in which it is used, this data may be linked to other types data which could potentially be used to identify an individual[316]. The same could also be said for thermometers and vibration sensors, depending on how they are combined with other sensory data.

[316] Fuller, Daniel, Martine Shareck, and Kevin Stanley. "Ethical implications of location and accelerometer measurement in health research studies with mobile sensing devices." *Social Science & Medicine* 191 (2017): 84-88.

Another ethical issue that may be inherent in the use of some types of sensors is safety. Infrared, radar and sonar can be classified as active sensors (though infrared can also be passive), which means they send out a signal, light wavelength or electrons to bounce off a target. Depending on the frequency of the signal sent out, there may a safety risk to humans or animals. Low frequency sonar, for example, has been noted as potentially dangerous for marine life and humans in the water.[317] Whilst infrared is not dangerous generally speaking, it could be if highly concentrated in a narrow beam.

A final set of issues with robot sensors relate to reliability and error. Sensors are used for measuring, and measurements are prone to errors. This depends on how well the sensors are performing – this can be influenced by a range of factors. They can vary in sensitivity, and as such there can be discrepancies between a sensor's output and the true value. The sensitivity of sensors can also lead to discrimination - based on race or nationality, for example. In one case, certain soap dispensers have been reported as not registering darker skin tones[318].

Sensor errors may be systematic and caused by factors which can be modelled, or they may occur randomly. Errors in sensor measurement, whether they are systematic or random, are not ethically problematic in themselves; however, they may lead to a variety of undesirable outcomes which are. Erroneous outputs may compromise a human or robot's ability to make good decisions, which in term may lead to undesirable outcomes. A robot measuring radiation levels to test whether an area is safe for humans, for example, might give an erroneous reading due to a faulty sensor. This could lead to a human decision which puts their own safety at risk.

Robot actuation

Actuators are the means by which a robot performs actions in its environment. They convert energy into mechanical form to produce movement, sound, vibration, light or chemical reactions. Widely used movement actuators include electric motors that produce torque to rotate wheels or gears, and linear actuators that create motion in a straight line. However, robots can have a variety of other actuators as well, including speakers, displays, LEDs, lasers, and other different types of movement-producing actuators. By means of its (electro)mechanical actuators, a robot can drive its other mechanical components and achieve complex motions with multiple degrees of freedom that are useful for object manipulation and locomotion. The "hand" of a robot is usually referred to as an (end) effector, while the "arm" is referred to as a manipulator. Robotic motion is studied in the fields of robot kinematics and robot dynamics. Robot kinematics is the study of the geometry of motion of a robot's mechanical parts, and robot dynamics is the study of the forces that are responsible for this motion.

Robots can be fitted with a wide variety of actuators. Some common robot actuators include:

- *Electric motors.* Electric motors convert electrical energy into mechanical energy. Magnetism forms the basis of their operation: one that is permanent and one electromagnet. Common types of electric motors for robotics include stepper motors, AC motors and DC motors.
- *Linear actuators.* Linear actuators create motion in a straight line (contrasting with the usual circular motion of an electric motor).
- *Piezoelectric motors.* Piezoelectric motors use a ceramic element which changes shape when an electric field is applied. This change produces a deformation or vibration which produces

[317] Parsons, E. C. M., Sarah J. Dolman, Andrew J. Wright, Naomi A. Rose, and W. C. G. Burns. "Navy sonar and cetaceans: Just how much does the gun need to smoke before we act?." *Marine pollution bulletin* 56, no. 7 (2008): 1248-1257.

[318] Aviva Rutkin, Digital discrimination, New Scientist, Volume 231, Issue 3084, 2016, Pages 18-19, ISSN 0262-4079, https://doi.org/10.1016/S0262-4079(16)31364-1.

and electrical charge. An electrical circuit produces acoustic or ultrasonic vibrations in the material, which then produces motion.

- *Speakers.* Speakers are made up of a cone, an iron coil, a magnet and housing. When electrical signals pass through the coil of the electromagnet, the direction of the magnetic field changes rapidly. This is picked up by the cone which amplifies the vibrations and pumps sound waves into the air.
- *LED displays.* An LED display (light-emitting diode display) uses a panel of LED lights as the light source. In robotics, they might be used to display information or even as an interaction medium between the robot and a human interlocutor.
- *Lasers.* Lasers (Light Amplification by Stimulated Emission of Radiation) produce a narrow beam of light in which all the wavelengths are lined up in phase. They can travel long distances and focus on very small spots.
- *Pneumatic artificial muscles.* Pneumatic artificial muscles are made mainly of a flexible and inflatable membrane. It has become popular in robotics due to its low weight and its compliant behaviour due to the compressibility of air.
- *Electroactive polymers.* Electroactive polymers are often referred to as artificial muscles. They are polymers which change their shape or size when stimulated by an electric field.
- *Biological.* Biological actuators are not 'biological' in a literal sense but generate movement similar to the musculature of a human being. This is part of a growing trend to make robots softer and safer.

The use some of these types of actuators may give rise to specific ethical issues, which include concerns about safety, health and bodily harm. With many types of motors, safety is a common concern. Some disadvantages of electrical motors, for example, include the possibility of overheating in static environments (in the presence of gravity). They may become an ignition source for fires. The presence of high-energy magnetic fields and high ferromagnetic forces of attraction may also pose a direct danger to health (to people with pacemakers, for example). Electrical based actuators may therefore pose a safety risk to people.

Safety concerns may also arise with actuators that produce sound or light. The usage of speakers at a high amplitude may cause bodily harm to humans through ear damage,[319] and general noise pollution is an environmental harm. For LED displays, there have been some investigations into potential bodily health risks, for example: the effect on those with photosensitive epilepsy; retinal damage; stress and annoyance and disruption of circadian rhythms.[320] Despite these concerns, there seems to be no direct adverse health risks, although there is the possibility of some discomfort to the eyes when exposed to blue light – particularly children.[321] There is some evidence of circadian rhythms being disturbed, although it is not clear if this leads to adverse health effects.[322]

[319] Passchier-Vermeer, Willy, and Wim F. Passchier. "Noise exposure and public health." *Environmental health perspectives* 108, no. suppl 1 (2000): 123-131.

[320] Oda, Joanna. Fong, Daniel. Zitouni, Abderrachid and Kosatsky, Tom. "Health Effects of Large LED Screens on Local Residents." National Collaborating Centre for Environmental Health. http://www.ncceh.ca/documents/practice-scenario/health-effects-large-led-screens-local-residents (retrieved 01/06/2019)

[321] Ibid.

[322] Ibid.

Ethical issues related to disability are also worth noting here. Individuals with epilepsy, for example, may have seizures provoked by the flicker frequency of screens[323]. More broadly speaking, designs of actuators may take it for granted that individuals have full vision, mobility, and hearing, thus forming a feedback loop of discrimination based on who can actually interact with the technology. This issue has been noted by authors who highlight the need for inclusive design[324].

Other actuators may pose safety risks if they are used improperly. Improper use of lasers can cause serious bodily harm to humans through thermal, acoustical and biochemical processes. These may range from mild skin burns to irreversible injuries. Artificial muscles used properly are also generally seen as safe. Pneumatic artificial muscles, for example, have been identified as non-hazardous as long as an innocuous gas is used in their operation.[325]

Now, let us turn to a specific category of actuator systems that enable locomotive capability in robots. Locomotion is a subfield that studies the various methods that robots use to transport themselves from place to place. This involves the design of both mechanical systems and control systems. There are numerous methods of robot locomotion. Some of these include:

- *Walking*. In contrast to wheeled motion, walking robots simulate human or animal motion. One of the main advantages of walking for a robot is the ability to negotiate inconsistencies in terrain.
- *Rolling*. In contrast to walking robots, which lose energy at heel strike when they touch the ground, rolling robots are the most efficient means of locomotion. Most rolling mobile robots will have four wheels or a number of continuous tracks.
- *Swimming*. Swimming robots may range from autonomous underwater vehicles (AUVs) which travel underwater without human input, or they may be bionic robots which have the shape and locomotion of a living fish.
- *Flying.* Flying robots are seen as particularly useful in surveying land, whether to map an area or on a search and rescue mission. Amongst the most popular types of flying robots are drones.

The actuators that enable these modes of locomotion may give rise to specific ethical issues, including concerns about safety, privacy and psychological harm. Many of the aforementioned locomotive methods may pose a safety risk to humans. Walking and rolling robots run the risk of bumping into humans; or of running into objects which may then become hazardous. The degree of this risk is dependent on a number of factors, including the size and speed of the robot itself. Flying robots, such as drones, present a unique risk of crashing from above, which has been noted as one of their most persistent safety issues.[326]

Robots of varying size and mobility could infringe on the privacy of others precisely because they can move around – perhaps into the private personal space of individuals. This risk of privacy infringement may be exacerbated depending on what kind of equipment the robots carrying and the kind of data it

[323] Ricci, Stefano, Federico Vigevano, Mario Manfredi, and Dorothée G. A. Kasteleijn-Nolst Trenité. 1998. 'Epilepsy Provoked by Television and Video Games, Safety of 100-Hz Screens'. *Neurology* 50 (3): 790. https://doi.org/10.1212/WNL.50.3.790.

[324] Abascal, Julio, and Colette Nicolle. "Why inclusive design guidelines?." In *Inclusive Design Guidelines for HCI*, pp. 21-32. CRC Press, 2001.

[325] Daerden, Frank, and Dirk Lefeber. "Pneumatic artificial muscles: actuators for robotics and automation." *European journal of mechanical and environmental engineering* 47, no. 1 (2002): 11-21.

[326] Custers, Bart. "Drones Here, There and Everywhere Introduction and Overview." In *The Future of Drone Use*, pp. 3-20. TMC Asser Press, The Hague, 2016.

is gathering. For example, a mobile robot that is carrying a camera poses an even greater risk to bodily, informational and relational privacy.

The previous point about privacy could be extended to a more general concern about psychological impacts of mobile robots. The mere presence of a mobile robot may be alarming for many people, particularly when the purpose of the robot is ambiguous to bystanders. This effect may be more acute in political contexts where dramatically different power dynamics exist. It has been noted, for example, that drones deployed in slums in East Africa instilled a fear of expropriation in some residents.[327] The visibility of this flying robot may therefore cause varying degrees of psychological harm.

Finally, let us now turn to a specific category of actuator systems that enable object manipulation. These are called *effectors* or *manipulators*. We can distinguish the following types:

- *Mechanical grippers* on a robot can grasp objects with mechanically operated fingers. They can be classified as electric or pneumatic grippers.
- *A vacuum gripper uses* a suction cup connected to a vacuum source to lift and move objects and are most effective when the object being gripped is smooth, flat and clean. They are commonly used in heavy industries.
- *Magnetic grippers* (classified as electromagnets or permanent) are most commonly used in a robotics for gripping ferrous materials. Electromagnets use a DC power unit and a controller unit for handling materials.

With regard to potential ethical issues, these types of grippers mainly have the give rise to concerns about safety and bodily harm. The environment in which the gripper is used will influence grip selection and safety considerations. For example, in the food and pharmaceutical industries, hydraulic actuated grippers are forbidden due to a risk of oil leakage and contamination. Vacuum grippers can create turbulent airflow and are thus not recommended in cleanroom industries. Special considerations must also be taken into account for the gripper's safe usage when used in toxic and corrosive environments.

Similar types of risks beset each type of gripper, though to varying degrees. One danger is that the work part that is being gripped is at risk of slipping out when the gripper is moving quickly, thus posing a risk of bodily harm to humans. Conversely, if the force applied by the gripper is too strong, this may cause bodily harm to a human if they are in contact with one another.

Robot control

The mechanical structures of robots must be controlled to enable them to perform tasks. Robot control systems take sensor data as input and calculate the appropriate signals to be sent to the actuators. These systems use techniques from (robot) control theory and can range in complexity. At a reactive level, they may translate raw sensor information into actuator commands in a relatively quick and simple fashion. However, at longer time scales or with more sophisticated tasks, they may need to use artificial intelligence and reason with cognitive models, which are intended to represent the robot, its environment, and the interactions between the two. Furthermore, robots may use pattern recognition and computer vision to track objects, techniques in robotic mapping to build maps of the world and localize themselves within these maps, and techniques in motion planning to figure out how they should move efficiently without hitting obstacles or falling over.

[327] Gevaert, Caroline, Richard Sliuzas, Claudio Persello, and George Vosselman. "Evaluating the societal impact of using drones to support urban upgrading projects." *ISPRS international journal of geo-information* 7, no. 3 (2018): 91.

The control system determines the robot's capacity for autonomous behaviour. Autonomy here can be defined as the capacity to operate in a real-world environment without external control. Robots can range in autonomy from fully autonomous to semi-autonomous. Fully autonomous or semi-autonomous behaviour in robots can range from basic to very sophisticated.

There may be four levels of designed autonomy:

- *Direct control.* This is a system that is unable to interact with and respond to its environment without human control.
- *Supervision.* Here, the robot selects and carries out options. The human monitors the system and intervenes if needed.
- *Semi-autonomy.* If a robot is semi-autonomous, it can be largely tele-operated, or be attached to and directly operated by the human body.
- *Autonomy.* These are robots which perform behaviours or tasks with a high degree of autonomy.

In terms of potential ethical issues, the capacity of robots for autonomous behaviour gives rise to concerns about safety, responsibility and accountability, transparency, privacy and discrimination. With a greater degree of autonomy comes greater safety risks, with researchers emphasizing that an autonomous robot requires high-precision data and quick reaction times in order to work safely around humans.[328,329] In the supervised and semi-autonomous systems, the need to hand off control from robot to human at various points of operation present challenges with safety implications.[330] Some of these include the need to decide what kind of situation requires a handoff; designing the ease of a handoff without significant disruption to functionality; and the need to avoid unwarranted human habituation to automatic controls (if a human is asleep at the point of hand-off, for example).[331] The so-called 'neglect curve' describes the relationship between user attention and robot autonomy, with the robot becoming less effective the more it is neglected and as the number of tasks increases in complexity.[332]

Related to the previous point and as has been noted in other sections, robot autonomy raises significant concerns about responsibility and accountability. Ethical and legal challenges present themselves in cases where a robot of semi- or full autonomy harms a human. It is not always clear in these cases where responsibility lies and who exactly should be held accountable. The increased autonomy of robots has the potential to change the human-robot relationship, with implications for the moral responsibility of the robot, safety regulations and design strategies.[333]

Issues of responsibility and accountability are closely linked to concerns over autonomy; namely, an increase in robot autonomy has engendered fears of a concurrent loss of autonomy on the part of

[328] Giuliani, Manuel, Claus Lenz, Thomas Müller, Markus Rickert, and Alois Knoll. "Design principles for safety in human-robot interaction." *International Journal of Social Robotics* 2, no. 3 (2010): 253-274.

[329] Kulić, Dana, and Elizabeth Croft. "Pre-collision safety strategies for human-robot interaction." *Autonomous Robots*22, no. 2 (2007): 149-164.

[330] Riek, Laurel, and Don Howard. "A code of ethics for the human-robot interaction profession." *Proceedings of We Robot* (2014).

[331] Ibid.

[332] Goodrich, Michael A., Dan R. Olsen, Jacob W. Crandall, and Thomas J. Palmer. "Experiments in adjustable autonomy." In *Proceedings of IJCAI Workshop on autonomy, delegation and control: interacting with intelligent agents*, pp. 1624-1629. Seattle, WA: American Association for Artificial Intelligence Press, 2001.

[333] Çürüklü, Baran, Dodig-Crnkovic, Gordana, & Akan, Batu, "Towards Industrial Robots with Human-like Moral Responsibilities", *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference,* April 2010.

humans – these relate to wider concerns, ranging from the impact on human dignity to the possibility that too much robot autonomy will lead to them 'taking over'.

Issues of responsibility and accountability also relate to the problem of transparency and the need to keep humans 'in-the-loop'. The decisions taken by a semi- and fully autonomous robot cannot be so technical and obscure that they are unintelligible to human operators. Biases may exist but be difficult to identify as it will not be clear how much weight is being given to each variable. As noted in the discussion on machine learning, unexplainable decision-making on the part of a robot goes against the 'need for explanation' and trust.[334]

Autonomy in robots typically require large amounts of data collection and, depending on the application, storage and usage of this data may infringe on the privacy of individuals. For semi- and fully autonomous robots, sensory data is relied upon in order for the robot to perform adequately. It is not always clear when this data collection is necessary for the robot's functionality or if it is extraneous. Who has access to this data also raises significant privacy concerns.

Let us now turn to the ethical issues in relation to some of the (what might be described as) "traditional" approaches in robot control (as contrasted with the more novel *robot learning approaches* that are discussed next). We can identify at least the following "traditional" approaches:

- *Robot learning.* Robot learning is a subfield that combines machine learning and robotics. It studies techniques that allow robots to acquire new skills or adapt to their environment through application of learning algorithms and/or neural networks.
- *Robotic mapping and motion planning*. A map is used for robots to localize themselves and for long term planning. The map may be known beforehand or generated during movement through the environment. In order to navigate the environment and avoid obstacles the robot needs a motion plan to move from an initial pose to a desired pose.
- *Adaptive control.* An adaptive control system is one that utilises a feedback control system in order to adjust its characteristics to parameters which are changeable or are uncertain at the start.

Here, we can mainly identify safety and reliability as a potential ethical issue. The above control approaches each raise concerns over reliability on several levels. The sensory data gathered by the robot must be accurate, when generating maps of an environment, for example. The algorithms used to process the data must be reliable such that the robot can adapt effectively to its environment. The efficiency on these processes will have a bearing on how safely the robot is able to operate, for example, whether it is able to successfully avoid hitting a human or an obstacle which may become hazardous to humans.

Now moving to *robot learning approaches*, we can identify the approaches:

- *Cognitive robotics.* These are robots which can learn from experience, from instructors, or on their own, and thereby develop the ability to effectively deal with their environment and react appropriately in real-world situations. This approach borrows from animal cognition models rather than more traditional artificial intelligence techniques.
- *Developmental robotics.* This describes an interdisciplinary subfield in robotics which studies the developmental mechanisms, architectures and constraints that allow for lifelong and open-ended acquisition of new skills and knowledge in robots. The approach aims to model increasingly complex cognitive processes in natural and artificial systems.

[334] Retrieved from: https://www.sophos.com/fr-fr/medialibrary/PDFs/other/GDPR-Pros-and-Cons.ashx

- *Evolutionary robotics.* This approach uses Darwinian principles of evolution to computer-simulate intelligent, autonomous robots with particular traits and unique skills. The best or fittest of these robots are iteratively selected and used as a basis for further diversification. Robots are treated here as organisms which can function independently of humans.
- *Behaviour-based robotics.* This subfield aims at creating robots that are capable of exhibiting complex-appearing behaviours despite having little internal variable state to model its immediate environment. The robots require no pre-set calculations to deal with a situation, and they are reactive in that they can correct their actions directly via sensory-motor links.

These robot learning approaches may give rise to ethical issues in relation to the loss of human control, responsibility and accountability, and justice and fairness. As noted in section 5.2.3, certain approaches to robot learning raise concerns about a loss of control on the part of humans. Evolutionary robotics, for example, poses a risk of allowing robots to develop beyond a point of human understanding and control. This is linked to concerns around transparency, with the possibility of robot motivations and decision-making becoming increasingly opaque and unpredictable, or even becoming biased, prejudiced and discriminatory whilst taken to be objective and neutral.

This previous point is again closely related to issues around responsibility and accountability, with more sophisticated robot learning methods potentially clouding the issue of who is responsible for the actions of a robot it its decision-making is the outcome of an increasingly opaque series of developmental iterations.

As has been noted in section 5.1.3 regarding algorithms, robot learning is built upon statistical models from data sets to predict future behaviour. Such predictions may be based on criteria which carries a bias against certain groups. This applies also to missing values and mistakes in the data. All of these issues can of course be exacerbated by the robot's decision-making being 'black-boxed'.

Now, let us move to a final aspect robot control we wish to discuss here, namely the emerging field of *cloud robotics*. Cloud robotics utilise cloud technologies centred on the convergence of information and communication infrastructures and shared services in the development of robotic systems. When connected to data centres in the cloud, robots can benefit from these centres' powerful (and relatively inexpensive) storage, computation and communication resources in the processing of data and the exchange of information with other robots.

Information that is stored and transmitted via the cloud is potentially at risk of hacking. The sensitivity of data stored in the cloud must therefore be taken into account, as the informational privacy of people may be at risk. The autonomy of robots also raises concerns about whether data is collected "incessantly".[335] Usage of cloud communication may have a significant impact on the types of information that are appropriate to reveal, share or transfer.[336]

5.2.3. Ethical issues with regard to general implications and risks

In this subsection, we describe the main ethical issues with regard to the general implications and risks of robotics technology. For each ethical principle and type of harm that we have identified as being implicated in any potential negative consequences of the development and use of robotics technology, we detail the ways in which harm can potentially occur. We focus on *loss of control*, *autonomy*, *privacy*,

[335] Pagallo, Ugo. "Robots in the cloud with privacy: A new threat to data protection?." *Computer Law & Security Review* 29, no. 5 (2013): 501-508.
[336] Ibid.

safety and security, *dual use and misuse*, *mass unemployment*, *human obsolescence*, *human mistreatment*, *robot rights*, and *responsibility and accountability*, respectively.

Loss of control

Human controllers may lose their grip on robotic actions by way of robot evolution. This ethical concern focuses the wisdom of creating robots that can grow and evolve beyond human understanding and control. Especially regarding the development of biological robots (see subsection 6.2.7), this is one of the biggest concerns behind creating self-sustaining and evolving robots is that they one day may surpass human understanding and control. As these types of robots would be very novel entities to humankind, their motivations, decisions, and actions would likely be opaque, leading to high degrees of unpredictability. When thinking of more present applications, unmanned vehicles, and military applications are particularly concerning as they have the means to cause significant amounts of death and destruction with incohesive policies and features to remedy unintended actions. This concern is always worth considering at every advancement of robots in any field as it would prove difficult to regain control once lost.[337,338,339,340]

Autonomy

Humans may become fully dependent on robots and may be incapable of survival without their aid. It is not so difficult to see how dependent human beings are on preceding technology, like electricity, running water, internet, telecommunications, automobiles, et cetera. This idea is particularly troublesome as humans are *already* very dependent upon various technologies and technological infrastructures, and if electric grids would somehow go dark, humankind would be in a large amount of trouble very quickly. It is uncertain how much robots would really add to this dilemma, or if it would add to the loss of human independence significantly more than any other technological advancement. In fact, if some of the environmental and maintenance robots are successful, it may help humans become more sustainable if robots are seen not as a fix, but as a redirection for the human community. It is pertinent to be mindful if one is creating robots that enable human self-sufficiency or are being used as an excuse not to change harmful human practices.[341,342] Further, at each increase of automatization, decisions and the power to decide, however incremental, is being taken from human beings. At some point, there may be a threshold in which so much decision-making power has been allocated to robots, that humans are unable to make certain types of decisions due to black-boxing of necessary information.

Privacy

Humans may no longer be able to expect privacy, as it is always possible that the robot may be collecting data and humans do not know what, where, or when. Privacy concerns remain a top ethical dilemma among all types of innovative technologies, and robots are no exception. The more sensory

[337] Torresen, Jim, "A Review of Future and Ethical Perspectives of Robotics and AI", *Frontiers in Robotics and AI: Evolutionary Robots*, January 2018.

[338] Hulme, David, "Rogue Robots", *Vision Insight: Global Threats*, August 2018.

[339] Kulkarni, Anagha, Chakraborti, Tathagata & Zha, Yantian et al., "Explicable Robot Planning as Minimized Distance from Expected Behavior", *Cornell University,* July 2018.

[340] Meinecke, Lisa & Voss, Laura, "I Robot, You Unemployed: Robotics in Science Fiction and Media Discourse", *Schafft Wissen: Gemeinsames und Geteiltes Wissen in Wissenschaft und Technik*, pp.203-215, October 2016.

[341] Torresen, 2018, op cit.

[342] Greenbaum, Dov, "Ethical, Legal and Social Concerns Relating to Exoskeletons", *Computers and Society* 45(3), September 2015.

data the robot relies upon to function, the more data it is going to need to be constantly collecting to ensure adequate performance. Whether this data be limited to a need-to-function basis, or additional data is being collected, remains unclear to users. Further, what the data is being used for and who has access to it and control over it leaves much room for ethical input. The more advanced robots become, the clearer the paramount nature of privacy-oriented questions will be, as the roles assigned to robots will heavily depend on the level of trust that can be assigned to them. If the potential for robots reporting confidential information and intimate interactions back to their companies for targeted advertising and analytics are too high, the growth of robots and their uses will be stunted. Even if individuals are willing to sacrifice some privacy for the sake of convenience, it is likely there will be a point of no return to where many robots will only be seen as advanced surveillance devices and not as mere machines or (for some robots) relational Others.[343,344,345] The side-effect of this being an increase of social paranoia and a “chilling effect” on society as it is no longer apparent who or what may or may not be observing human behaviour.

Safety and security

Robots could cause a great deal of harm if they suffer a computer security breach or have design flaws. In cases where robots have a large amount of responsibility for humans and trust, for example, hospitals, military contexts, elderly or child care, the prospect of an individual gaining unauthorized access to a robot in these scenarios would be a profound concern, especially if the human interactors are not aware of there being a security breach or unable to regain control of the robot. Accordingly, it is incredibly important that security measures, parameters, and safeguards are implemented and followed that evolve with the robot. If security and safety designs, policies, or procedures begin to lag behind the robot’s societal responsibilities and capabilities, the potential risks are great.[346,347] Further, even while using robots appropriately and following design protocols, there is the potential for robots to malfunction or function unexpectedly that may potentially lead to human harm. Consequences of machine malfunctions during approved use may be enough to kill the technology’s implementation in near-future applications. Additionally, if robots are particularly susceptible to security breaches or are sneakily reporting data back to its corporate creators for use of advertising and analytics, sensitive fields, like healthcare, may want to carefully consider if robots are the best fit for them. This also threatens the already-fragile trust of robots at present.

Dual use and misuse

Robots may be used in ways unintended by their creators. This set of ethical dilemmas is really focused upon during the design and creation part of robots, as designers and engineers have the largest hand in eliminating potentials for misuse and dual use. Unfortunately, even when trying to make design choices that eliminate these possibilities, it is impossible to control for everything. As such, it still stands that sex robots could be used for spying or a food delivery robot could be used to breach buildings.[348] Or friendly security robots could be modified into something more nefarious. There are

[343] Gonzalez-Fierro, Miguel, “10 Ethical Issues of Artificial Intelligence and Robotics”, Github, April 2018.
[344] Rueben, Matthew, Bernieri, Frank & Grimm, Cindy et al., “Framing Effects on Privacy Concerns about a Home Telepresence Robot”, *2017 IEEE International Conference on Human-Robot Interaction*, March 2017.
[345] Kirschgens, Laura, Ugarte, Irati & Uriarte, Endika et al., “Robot Hazards: From Safety to Security”, Whitepaper, 2018.
[346] Simon, Matt, “The Serious Security Problem Looming Over Robotics”, *Wired Science*, August 2018.
[347] Booth, Serena, Tompkin, James & Pfister, Hanspeter et al., “Piggybacking Robots: Human-Robot Overtrust in University Dormitory Security”, *Human Robot Interaction*, March 2017.
[348] Booth, 2017, op. cit.

seemingly few regulations and rules that address the issue of robot modification and misuse, in a way it is understandable. If the regulations lean too heavily towards the favour of non-modifiable robots, it might be difficult for individuals to perform their own maintenance, repairs, or experiments on their own devices—much like cellular devices of present times. However, with no regulations at all, leaves the question too open-ended, and it may be likely problems will occur similarly to the ethics surrounding 3D printed weapons. For this area of ethics, it is difficult to find a middle ground between beneficial modification allowances and misuse.

Mass unemployment

There is still much uncertainty about the impact of robots in terms of unemployment. Robots *may* take over human jobs that cause unemployment rates to rise, but already present issues of exacerbated socio-economic inequality. While it is always important to be mindful of a robot's impact on the labour market and labourers themselves, many of the concerns tend to be out of proportion to the scale and speed of automation. Further, as more problems begin to surface with fully automated business strategies, many companies are looking towards collaborative robotic solutions. These solutions utilise robots for monotonous or dangerous tasks, while human labourers work with the robots on more complicated tasks. Not only bumping up the quality and speed of labour, but also easing the burden of these tasks on human workers. While this may lead to a large amount of job layoffs from these positions, recent studies suggest that the human job market will flow into the areas required to keep these robots up-and-running, and to perform more difficult tasks that robots are not yet capable of achieving. Now, the more concerning area of this rerouting, and one that does not generate as much attention, is the facilitation and worsening of existing socio-economic class stratifications and power-relations. Further, keeping a sharp eye on worker conditions and ensuring that the workload and expectations of labourers is not increased without adequate compensation and training. The jobs themselves do not seem to be as problematic as the societal fallout from such a change.[349,350] (For thorough description of mass unemployment issues as they relate to AI, but also to robotics, see the previous discussion on "Responsibility and accountability" of subsection 5.1.3.)

Human obsolescence

Over the long term, we *may* arrive at a future where robots have become so superior to human beings so that humans will lose their place and purpose. This concern is more often formulated in media and science fiction as "robots taking over the world" and is a concern that is often a combination of other human dignity concerns like: "loss of control", "human mistreatment", and "human obsolescence". Most of these debates and discussions are on many far-off iterations of humanoid androids or robots, but it still stands worth mentioning as these moral and existential concerns will still guide the creation, policy, and research surrounding robots and their advancements, even if they are unwarranted at present. Using ethics to not only help individuals come to terms with robotic others, but also to come to terms with and understand that the meaning of 'being human' will also change in a new technological era. The importance of ethics at this time will be as important for guiding the development of humans as it will robots—as many individuals will likely turn to the arts and humanities

[349] Barlow, Rich, "Economist Predicts Job Loss to Machines, but Sees Long-Term Hope", *Psys.org Robotics*, March 2018.

[350] Vincent, James, "AI and Robots will Destroy Fewer Jobs than Previous Feared, Says New OECD Report: But the Impact Will Still be Significant, Increasing Societal Division Between the Rich and the Poor", *The Verge*, April 2018.

for guidance when they feel a loss of identity is imminent, as humankind has done in the past with cultural transitions.[351,352]

Human mistreatment

If the development of robots goes too far, they may evolve to treat humans poorly or harm us. Especially with high risks of inequality and discrimination being learned by robots, it is critical that the algorithms robots are using for decisions and the sensory information gleaned by robots are being carefully monitored for biases. To prevent such situations, some authors call for more transparency in machine decision-making processes and starting data points. While this may not completely fix data biases and discriminatory decisions, it would allow for more participation and monitoring for these problems than black-boxing this information would. Furthermore, other researchers suggest setting hard parameters on how robots are permitted to interact with humans, e.g. not killing human beings or no robots allowed in law enforcement. Ethics stands to have much to offer in how this area will develop, and it is important that these frameworks are decided upon and implemented before the robots are given free rein in their roles.[353,354,355]

Robot rights

Undoubtedly one of the most complicated issues in robot ethics, the question of robot moral standing respective to humans and animals is one that generates much debate. Questions on whether moral responsibilities, duties, and treatment are owed to robots, and, if so, to which types of robots and what those duties, responsibilities, and treatment entail, are important. And not only for the sake of the robots, but the ways in which humans treat robots, especially those designed specifically to imitate human beings, may reveal some uncomfortable truths about those human beings that need to be addressed. While it may not be pragmatic to jump to personhood status for, even some, robots like Saudi Arabia has decided,[356] there is something to be said for epistemic caution when approaching the idea of robot rights. At the very least, prohibiting individuals from physically attacking robots, preventing them from performing their assigned roles, or interacting with them maliciously (i.e., bullying) may prove beneficial to paving the way for robotic community members of the future.[357,358,359,360]

Responsibility and accountability

If robots cause harm or destruction, who is responsible for reparations? One of the most frequently discussed question, both in the academic spheres and in the media, is that of robot responsibility and accountability. Especially pertinent in ongoing discussions about self-driving (or “autonomous”) vehicles, who is to blame when the machine malfunctions? The more complex and black-boxed a machine’s decision-making models and processes are, the more difficult it becomes to determine who

[351] Torresen, 2018, op cit.
[352] Mussolum, Erin, “How Art Shapes Identity”, Trinity Western University, October 2007.
[353] Hulme, 2018, op. cit.
[354] Kulkarni, 2018, op. cit.
[355] Mulligan, Christina, “Revenge Against Robots”, Brooklyn Law School, 2017.
[356] https://www.pri.org/stories/2017-11-01/saudi-arabia-has-new-citizen-sophia-robot-what-does-even-mean
[357] Snow, Jackie, “A Robot’s Biggest Challenge? Teenage Bullies”, *Technology Review: Intelligent Machines*, March 2018.
[358] Ackerman, Evan, “Robotic Tortoise Helps Kids to Learn That Robot Abuse is a Bad Thing”, *Spectrum IEEE*, March 14.
[359] Gunkel, David, Robot Rights, The MIT Press, 2018.
[360] Gordon, John-Stewart, “What do we owe to intelligent Robots?,” *AI & Society*, 2018.

or what is responsible. This is particularly important when it comes to determining how to compensate damages and harm done by robots— if a self-driving vehicle crashes and kills its driver due to a faulty decision-making protocol, is it the company responsible for the malfunction? The QA board for not catching the error before deployment? The driver for not monitoring driving conditions? All of these entities? None of them? Before robotics hit ubiquity, it is critical to establish chains of responsibility for these technologies and formulate legal and regulatory policies to account for non-human decision-makers.[361,362] (For a more thorough description of responsibility and accountability issues, see the discussion on "Responsibility and accountability" of subsection 5.1.3.)

[361] Gonzalez-Fierro, 2018, op. cit.

[362] Booth, 2017, op. cit.

6. Ethical analysis: Ethical issues with AI and robotics products

In this section, we identify and describe the main ethical issues with regard to artificial intelligence and robotics technology *products and procedures*. As stated in the methods section, in this ethical analysis, we follow the *Anticipatory Technology Ethics* approach developed by Brey (2012).[363] Having focused on the technology level in the previous section, we now turn our attention to the *artefact level* (or *product level*) of the approach's three-level system of ethical analysis.

Our objects of analysis at this level consist of technological artefacts (i.e., physical products) and technological procedures (i.e., functional procedures developed within the field) that are being developed on the basis of AI and robotics technology for use outside of these fields. Thus, in this section, we discuss the ethical issues that are either inherent in or may occur across a wide range of applications of such products of AI technology as *intelligent agents*, and *computer vision systems*, as well as such products of robotics technology as *social robots* and *unmanned aerial vehicles*.

In this section, we again focus on both present issues and issues that may occur between now and 20 years into the future. Most of our analysis in this section is based off of an extensive analysis of the academic and popular literature on ethical issues in AI and robotics products and procedures. Additionally, we have made use of the results of our SIENNA expert workshops and expert interviews, and we have on occasion used ethical checklists to conduct our own analysis in areas where the literature was sparse.

This section is structured as follows. Subsections 6.1 and 6.2 describe the ethical issues inherent in AI products and in robotics products, respectively. Each of these subsections discusses a range of present and potential future classes of products and procedures, their properties, and the potential ethical issues that are may occur in relation to them.

6.1. Ethical issues with AI products

This subsection identifies and describes the potential ethical issues that are either inherent in, or may occur across a wide range of applications of, important kinds of AI products. It discusses, in turn, the issues for *intelligent agents* (subsection 6.1.1), *knowledge-based systems* (subsection 6.1.2), *computer vision systems* (subsection 6.1.3), *natural language processing systems* (subsection 6.1.4), *affective computing systems* (subsection 6.1.5), *(big) data analytics systems* (subsection 6.1.6), and *embedded AI and Internet of Things* (subsection 6.1.7). Table 8 below lists the most important ethical issues that have been identified for each of these types of AI products.

Type of product	Ethical issues	
Intelligent agents	**- Autonomy and freedom** **- Privacy** **- Responsibility and accountability** **- Safety** **- Security**	**- Trust** **- Human dignity** **- Diminishing of social interaction** **- Social de-skilling**

[363] Brey, P.A.E., 2012, op cit.

Knowledge-based systems	**- Bias** **- Accuracy**	**- Unpredictable outcomes** **- Security**
Computer vision systems	**- Security** **- Privacy**	**- Accuracy**
Natural language processing systems	**- Privacy** **- Bias and discrimination** **- Transparency**	**- Accuracy**
Affective computing systems	**- Privacy** **- Trust** **- Autonomy**	**- Potential for deception** **- Unwanted social bonding**
(Big) Data analytics systems	**- Privacy** **- Bias and discrimination**	**- Transparency** **- Responsibility and accountability**
Embedded AI and Internet of Things	**- Privacy** **- Security** **- Trust**	**- Autonomy and freedom** **- Responsibility and accountability**

Table 8: Overview of ethical issues with major types of AI products.

6.1.1. Intelligent agents

Please note that ethical issues with intelligent agents that qualify as robots are discussed under various categories in section 6.2 on robotics products and in section 7.2 on robotics applications.

Intelligent agents are autonomous, artificially created entities[364] that perceive their environment through sensors, act upon that environment using actuators, and direct their activity towards achieving goals (i.e., they are "rational" agents). Over the last few decades, AI technology has advanced to a level that has enabled billions of intelligent agents to do their work in people's smartphones, smart appliances, Internet search engines, self-driving cars, electronic markets, military equipment, care robots, et cetera. AI techniques and products such as machine learning and natural language processing (NLP) systems have allowed intelligent agents to better process user input, learn new skills, and make decisions based on large and difficult sets of parameters. The result of this is that they can decide, act, interact, and adapt autonomously in very complex and dynamic real-world environments, enabling us to let them drive our cars on public roads, make suggestions on which political party to vote for during elections, and be companions for our lonely grandparents in the nursing home.

A large variety of types of intelligent agents currently exists: intelligent assistants (e.g., in smartphones), customer service chatbots, virtual companions, and non-human players in videogames are some of the most familiar examples. Different types of intelligent agents can vary greatly in terms of their basic characteristics and ethically relevant dimensions: they may have (1) varying degrees of perceptibility for humans; (2) different levels of operational engagement with users; (3) different levels of authority or control with respect to the user's actions; (4) different kinds of embodiment; (5) different abilities in terms of social interaction with humans; (6) different capacities to learn new behaviour; and (7) different levels of interaction with humans, other agents and computer systems, amongst others. In what follows, we briefly describe each of these dimensions and identify the (potential) ethical issues raised by them.

[364] For the purposes of this section, we can consider them software programs.

To begin, intelligent agents can have different levels of perceptibility for humans. They can, for example, be designed as virtual agents with discrete sensors that engage in stealthy observation for security purposes. Agents' low perceptibility may raise the potential for privacy issues and have a chilling effect on people's behaviour.

Second, intelligent agents can have different levels of operational engagement with users. Agents might require higher or lower levels of user input, and they might provide users with information on a more frequent or less infrequent basis. High levels of interaction with users can be a distraction for users and may, in certain situations, present safety concerns. On the other hand, low levels of interaction can make agents less perceptible to users and, as such, may present privacy concerns.

Third (and somewhat related to the second dimension), intelligent agents can be designed to have different levels of authority or control with respect to the user's actions in their applications contexts. There are agents whose task it is to *assist* the user in or by carrying out certain actions; there are agents whose goals are to (actively) *persuade* the user to perform particular actions; and there are agents who are designed to (take) *control* (away from) the user under specific circumstances. Agents are currently already being used for a broad range of persuasive purposes. Although many of these are fairly innocuous (e.g., a fitness tracker persuading its user to run an extra kilometre), it is not hard to imagine questionable (e.g., persuasive agents making recommendations on, for example, who to vote for in elections, who to date, and what career choices to make) and malign (e.g., by being manipulative and coercive) interventions by intelligent agents.[365] Some recommender systems may attempt to "addict" users to certain types of "contents".[366] Ethical concerns about high levels of agent authority may relate to such values as autonomy, freedom, moral responsibility, human dignity, safety, and trust.

Fourth, intelligent agents can have different kinds and degrees of embodiment. Intelligent agent embodiment refers to the state of being constructed out of physical materials (*robotic embodiment*), *appearing* to be, but not actually being, constructed out of physical materials (*virtual embodiment*), or not being embodied. Experimental studies have shown that different kinds of (non-)embodiment give rise to different effects on users in terms of receptiveness for persuasion, performance, trust, and well-being.[367,368] Ethical concerns with regard to the nature and level of embodiment can relate to values such as trust, human dignity, privacy, and general well-being. For example, a decision to include in a health care setting intelligent agents that are embodied virtually rather than physically might have negative implications in terms of patients' trust and their experience of loneliness. On the other hand, including physically embodied agents in such a setting might lead to a reduction in experienced privacy.

Fifth, intelligent agents can have different abilities in terms of sociability. The concept of sociability here is related to that of embodiment, but distinct from it in that it focuses on agents' social *behaviour*. Intelligent agents can exhibit a wide range of social behaviour. In the context of robotics, Breazeal (2003) has defined four classes of social robots in terms of the complexity of their capacity for social interaction: *socially evocative robots*, *social interface robots*, *socially receptive robots*, and *sociable*

[365] Milano, Silvia, Mariarosaria Taddeo, and Luciano Floridi, "Recommender Systems and their Ethical Challenges," 2019. http://dx.doi.org/10.2139/ssrn.3378581

[366] Burr, Chistopher, Nello Cristianini, and James Ladyman, "An Analysis of the Interaction Between Intelligent Software Agents and Human Users," *Minds and Machines*, Vol. 28, No. 4, 2018, pp. 735–774.

[367] Li, Jamy, "The benefit of being physically present: A survey of experimental works comparing co-present robots, telepresent robots and virtual agents," *International Journal of Human-Computer Studies*, Vol. 77, 2015, pp. 23–37.

[368] Rickenberg, Raoul, and Byron Reeves, "The effects of animated characters on anxiety, task performance, and evaluations of user interfaces," *CHI '00: Proceedings of the SIGCHI conference on Human factors in computing systems*, New York, NY, USA: ACM Press, 2000, pp. 49–56.

robots.[369] It appears that a similar classification is possible for intelligent agents. Conversational agents, which possess an ability to understand natural language in speech or text, and converse with a human in a coherent way, would be a good example of a category where sociability is very high. The use of highly social agents in certain application contexts may generate significant ethical concerns relating to such values a privacy, trust (e.g., deception through simulated emotional responses), safety (e.g., potential unwanted distraction), and general well-being (e.g., potential reduction in human-to-human contact, loss of community, insufficient ability to recognise particular social sensitivities, social de-skilling).

Sixth, intelligent agents can differ in terms of their adaptivity. In the last decade, the use of machine learning techniques has been instrumental in the development of intelligent and adaptable intelligent agents. Machine learning techniques give agents the ability to learn new behaviour without this behaviour having to be explicitly programmed. Insofar as the intelligence and adaptivity of agents are based on machine learning techniques, there are increasing concerns about not having a rich explanatory and predictive account of the behaviour of these agents. This is true in particular for such currently widely researched AI techniques and methods as artificial neural networks, deep learning, and genetic algorithms. The ethical issues here may relate to moral responsibility and accountability (e.g., the responsibility ascription problem[370]), safety, trust, and justice (e.g., potential for algorithmic bias[371]), amongst other values.

Seventh, intelligent agents can have different levels of connectedness, meaning that they can have different levels interaction with humans, other agents, and computer systems outside the use context. At present, many agent systems are not smart enough to do all the necessary processing on their own. Consequently, they need to rely on cloud computing – on servers that are often located in a distant country – to parse user input and turn it into usable information. Agents may also be designed to communicate with other agents and humans so as to better serve their users. Such increased connectedness raises the potential for privacy (e.g., access and/or control over personal information) and security issues (e.g., hacking).

Besides the above dimensions, other ethically relevant aspects of intelligent agents include perception (i.e., the range and sensitivity of agents' senses), actuation (i.e., the range and effectiveness of agents' actions), and moral intelligence. The first and second of these may implicate significant privacy and safety issues, respectively. More complicated, however, is the ethical discussion on the application of "moral intelligence" in intelligent agents; for a detailed description of the ethical issues in this area, readers are referred to the part on "Machine ethics" in subsection 5.1.3 of this report.

[369] Breazeal, Cynthia, "Toward sociable robots," *Robotics and Autonomous Systems*, Vol. 42, 2003, pp. 167–175.
[370] See the part on "Responsibility and accountability" in subsection 5.1.3 of this report. The responsibility ascription problem refers to the problem of ascribing moral responsibility for the harmful consequences of an agent's self-learnt behaviour. In such cases, moral responsibility may not easily be ascribed to the system's designer or anyone else. Matthias, 2004, op. cit.
[371] See the part on "Justice and fairness" in subsection 5.1.3 of this report. Algorithmic bias can occur when the data that is being used to teach a machine-learning-based agent reflects the implicit values of humans involved in the collection, selection, or use of the data. Nissenbaum, Helen, "How computer systems embody values," *Computer*, Vol. 34, No. 3, 2001, pp. 120–119.

Finally, it deserves to be emphasised that whether any of the ethical issues that have been identified in relation to the aforementioned dimensions will occur in practice is in part dependent on the specifics of the application context.[372]

6.1.2. Knowledge-based systems

In the field of AI, *Knowledge Engineering* encompasses the construction, maintenance and application of *knowledge-based systems* (KBSs), and all related technical, scientific and social aspects. Knowledge-based systems are computer programs that use a knowledge base to draw inferences and solve complex problems. The earliest forms of KBS were *expert systems*: Computer systems which aim to mimic the decision-making capabilities otherwise performed by a human expert.[373] Expert systems were among the first manifestations of AI programs: Emerging during the 1970s, expert systems use a reasoning system to analyse large quantities of information, allowing them to produce new information which then can be applied to solve complex problems.

Although expert systems were the first knowledge-based systems, nowadays it is important to note that there is a significant difference between the two. Expert systems are defined by their function and task. They assist in a given task by applying expert knowledge and analysis to a given problem. KBSs, on the other hand, are defined by the architecture of the systems themselves: Instead of using procedural code—as "standard" databases do—a KBS explicitly represents its knowledge (see subsection 5.1.2 on knowledge representation and reasoning techniques). A KBS has (at least) two subsystems, which are also its defining features: A *knowledge base* and an *inference engine*. First, a knowledge base is a storage system which contains complex information about the world, commonly structured in a type of ontological model (ideally an object model supporting classes and instances). Contrary to a database, the available data in the knowledge base is structured by the inference engine according to its set of rules, which allow it to derive new facts from the dataset. Later types of architecture for KBS developed the possibility to allow the reasoning process to affect its own procedures using the inferences from the original reasoning parameters. As such, KBS evolved to not only apply themselves to solving a specific problem within a field, but also diagnose potential problems with its own reasoning on the subject.[374]

In addition to expert systems, KBSs also encompasses other forms of intelligent knowledge-based systems, including: logical operations controllers,[375] educational systems,[376] recommender systems,[377] and tools for data-mining,[378] knowledge management,[379] accounting,[380] computer system

[372] To illustrate this point, consider a highly sociable agent that is talkative and out-going. Such a quality may in many situations facilitate social interaction, which can be very helpful in for example a care setting with elderly persons in a nursing home. In another application, however, the sociality of such an agent *can* be detrimental. For example, a chatty agent in a car may distract the driver from the road – or perhaps, as suggested by Eriksson and Stanton (2016), it could be used as a co-driver helping the driver of a semi-autonomous vehicle to keep his or her attention on the road. Eriksson, Alexander, and Neville Stanton, "The chatty co-driver: A linguistics approach to human-automation-interaction," In: *Contemporary ergonomics and human factors 2016: Proceedings of the international conference on ergonomics & human factors*, 2016.

[373] Naser and Zaqout, "Knowledge-Based Systems That Determine the Appropriate Students Major", 26.

[374] Faniyi et al., "Architecting Self-Aware Software Systems".

[375] Nan, Khan, and Iqbal, "Real-Time Fault Diagnosis Using Knowledge-Based Expert System".

[376] Naser and Zaqout, "Knowledge-Based Systems That Determine the Appropriate Students Major".

[377] Tarus, Niu, and Mustafa, "Knowledge-Based Recommendation".

[378] Choudhary, Harding, and Tiwari, "Data Mining in Manufacturing".

[379] Dalkir, Knowledge Management in Theory and Practice, 217–244.

[380] Dillard and Yuthas, "Ethics Research in AIS".

diagnostics,[381] medical diagnostics,[382] computer design tools,[383] case-based reasoning[384] and knowledge retrieval (database retrieval systems & information retrieval systems, such as web search engines).[385] KBSs is thus a very broad field with a lot of different ethical implications related to the specific context to which a KBS is applied.

With the rise of KBSs has come the possibility of manipulating and controlling knowledge. In the broader sense of *knowledge management*, the sources of the initial data, ways of collecting knowledge, the design of the rules and algorithms of the inference engine, storage of data and its eventual distribution are all accompanied by potential ethical issues inherent to the system itself.[386] KBSs have the ability to create knowledge, yet it can also be omitted, suppressed, amplified, exaggerated, diminished, distorted or destroyed.[387] These problems can arise with and without the intentionality of the designer in regard of doing so. The core ethical issue with regard to KBSs as such is the manipulation of knowledge: In this context, the ethical issues relate primarily to how the architecture of KBSs influences the outcomes of the problem which is being investigated: How can forms of tacit discrimination or domination be prevented when implementing a system which cannot be held accountable as a thing in itself? If a KBS has unforeseen consequences, who is responsible for the implications? Another potential ethical issue concerns the implementation of KBS systems: How do we ensure they are applied for the "greater good", as opposed to potential hidden political or corporate agendas?[388]

A second set of ethical concerns emerge from the fact that KBSs can behaviourally adapt to the knowledge they themselves produce: What happens if we ethically design a KBS system which then adapts towards a more unethical approach? The nature of KBSs necessitates the gathering of information about the world, which has to be translated into the language of the knowledge base, and back into the real world after it has been manipulated.[389] Potential ethical dilemmas of this kind are related to the translation of knowledge.[390] For example, when data on citizens is gathered and analysed by a KBS, how much control do these citizens have over what parts of their identity are being captured, processed and stored? There is a potential for unfair exploitation, stigmatization, profiling or malicious application due to the fact that an identity has to be reduced to a set of variables and parameters when handled by a KBS. These are questions of autonomy, privacy, but also of ownership of identity, which are partially being negotiated between the control of humans over the KBS and the (semi-)autonomous adaptations the KBS itself deems necessary to succeed.[391] Furthermore, the difference between statistical significance of outcomes and practical significance is an often neglected subject in academic debate regarding information technology.[392] Considering that factor, it has to be accounted for that

[381] Hu, Schroeder, and Starr, "A Knowledge-Based Real-Time Diagnostic System for PLC Controlled Manufacturing Systems".
[382] Hayes-Roth and Jacobstein, "The State of Knowledge-Based Systems".
[383] Gennari et al., "The Evolution of Protégé".
[384] Aamodt and Plaza, "Case-Based Reasoning".
[385] Burke, "Knowledge-Based Recommender Systems".
[386] Abbasi, Sarker, and Chiang, "Big Data Research in Information Systems", 24.
[387] Land, Amjad, and Nolas, "Accountability and Ethics in Knowledge Management", 2.
[388] Bryant, "Knowledge Management — The Ethics of the Agora or the Mechanisms of the Market?"
[389] Abbasi, Sarker, and Chiang, "Big Data Research in Information Systems", 7.
[390] Akhavan et al., "Exploring the Relationship between Ethics, Knowledge Creation and Organizational Performance", 44.
[391] Mason, "Four Ethical Issues of the Information Age".
[392] Lin, Lucas, and Shmueli, "Research Commentary —Too Big to Fail", 9–10.

the values we neglect to inscribe in research and practice will also be practices that will not properly be accounted for in the design of KBSs.

This brings us to a third set of potential ethical implications, regarding accuracy and access. Accuracy should be understood as the measure of quality of the generated knowledge. Here we can distinguish two issues: unintended errors, and information which is intentionally misleading. In the first sense, quality control and ethical design will solve the majority of preventable issues.[393] The question here is therefore rather about who bears the responsibility for these quality measures. In the second sense, questions of accountability for accuracy as well as system integrity will become increasingly important. This brings us to the value of access: Access should be understood as two issues as well: In the first instance, who should have access to the KBS and its product, whom needs to have access for, e.g., oversight purposes? Who shouldn't have access due to the potential sensitivity of the gathered knowledge? The second instance is a matter of security, with questions regarding the distribution of access needing to be fairly distributed among stakeholders and interest groups. Both issues come down to matters of transparency: Important questions will be on the verifiability, authenticity and robustness of KBSs and their methodologies.

6.1.3. Computer vision systems

Computer vision systems developed alongside AI and continues to be one of its most significant applications. These systems interpret visual information to identify objects visible in an image or in video footage. Depending on their intended purpose, computer vision systems may implement one of several approaches to interpreting visual data: feature detection, recognition, and reconstruction.[394] Feature detection (sometimes referred to as feature extraction) applies algorithms directly to the image received by the system to identify characteristics of objects within it. For example, the edges of objects may be identified by searching for significant changes in brightness within the image.[395] Recognition attempts to identify objects within an image by searching for patterns.[396] Human faces have a regular arrangement of features (eyes, nose, mouth) that can be identified even though individual faces differ from one another. Finally, reconstruction is used to construct a geometric model within the computer vision system that represents the objects identified within the image.[397] This may be performed by analysing different images of the same objects to identify differences between them, and by identifying motion, lines, contours, and textures within the images.[398]

Computer vision systems have a variety of applications. The applications with the most significant ethical concerns are object detection, image classification and object recognition, and visual biometric systems (such as face, iris and fingerprint identification). Each of these applications raise concerns about safety, privacy and the expanded monitoring and surveillance capabilities that they offer for governments, employers, and individuals.

As mentioned above, objects may be identified by searching for patterns within images. People may be distinguished from other objects in video footage by searching for specific sets of features within an image.[399] This is important for computer vision systems in automated vehicles or autonomous

[393] Dillard and Yuthas, "Ethics Research in AIS", 10.
[394] Russell, S., and Norvig, P., *Artificial Intelligence: A Modern Approach,* 3rd ed., Essex, Pearson, 2016, p. 929.
[395] Ibid., p. 936.
[396] Ibid., p. 942.
[397] Ibid., p. 929.
[398] Ibid., pp. 947-957.
[399] Ibid., pp. 945-946.

robots that must navigate themselves around people. Failures in object detection will cause the autonomous robot or vehicle to crash into objects or people, and potentially cause damage or injury.

Object recognition systems are useful for automatically captioning photos, which may increase the accessibility of computers and social media to visual impaired users.[400] The accuracy of image classification and object recognition systems depends on whether the image data sets are fully representative of the objects they are intended to classify and recognise. Image classification systems may exhibit unintentional bias and reflect prejudices if they are trained with unrepresentative data. Google was forced to apologise in 2015 after its Photos service labelled photos of two African-Americans (a software developer and his friend) as 'gorillas'.[401] Gender bias may also appear in object recognition systems that are more likely to identify people shown in kitchens as women.[402]

Computer vision has greatly expanded the possibilities for using visual biometrics, which identify specific individuals through the visual recognition of an individual's characteristics, such as their face, their fingerprint, or the irises of their eyes.[403] They may be used to authenticate someone's claim to an identity or to identify people visible in pictures or video footage.

For authentication, such biometrics have significant practical benefits compared to other methods of establishing identification, such as ID cards, as they are intrinsic to the identified person, cannot be easily forged, and cannot be misplaced, stolen or shared with others.[404] However, they also impose additional difficulties for individuals whose physical appearance and attributes change through accident, illness, or choice. People with finger or eye injuries, for example, may no longer be identifiable via fingerprint or iris recognition. Changes to facial appearance may also prevent people from using facial recognition systems. Automated Gender Recognition (AGR) through facial recognition raises concerns about how it assumes that gender necessarily corresponds with sex.[405] For example, the ride-sharing platform Uber requires drivers using its platform to occasionally verify their identity by taking a photo of their face. Uber drivers who were transitioning to a different gender found themselves unable to verify their identity using this method as the facial recognition system no longer recognised them as the same person.[406]

The almost ubiquitous use of CCTV cameras in public areas by governments and law enforcement agencies means that individuals in most developed cities will appear in video footage. Similarly, video cameras may be used within stores to monitor customers and identify shoplifters. The automation of video surveillance made possible by computer vision technology has several potential benefits over having human operators observing video input. Automated video surveillance allows for more visual data to be processed, while human operators are likely to resort to social stereotypes to determine

[400] Wu, S., Wieland, Farivar, O., and Schiller, J., 'Automatic Alt-text: Computer-generated Image Descriptions for Blind Users on a Social Network Service', *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, Portland, Oregon, USA, pp. 1180-1192, February 25-March 1, 2017.

[401] Simonite, T., 'When It Comes to Gorillas, Google Photos Remains Blind', *Wired*, January 11, 2018. https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

[402] Simonite, T., 'Machines Taught by Photos Learn a Sexist View of Women', *Wired*, August 21, 2017. https://www.wired.com/story/machines-taught-by-photos-learn-a-sexist-view-of-women/

[403] Kroeker, K. L., 'Graphics and Security: Exploring Visual Biometrics', *IEEE Computer Graphics and Applications*, Vol. 22, No. 4, July 2002, pp. 16-21.

[404] Prabhakar, S., Pankanti, S., and Jain, A. K., 'Biometric Recognition: Security and Privacy Concerns'*, IEEE Security & Privacy*, Vol. 1, No. 2, pp. 33-42, March-April 2003.

[405] Keyes, O., 'The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition', *Proceeding of the ACM on Human-Computer Interaction*, Vol. 2, No. CSCW, Article 88, November 2018.

[406] Urbi, J., 'Some Transgender Drivers Are Being Kicked Off Uber's App', CNBC, August 13, 2018. https://www.cnbc.com/2018/08/08/transgender-uber-driver-suspended-tech-oversight-facial-recognition.html

who to focus on during surveillance.[407] However, these flaws may also affect automated surveillance. Prejudices against certain people may be reflected in the software incorporated into the system, as well as expectations about individual and group behaviour that are culturally specific and do not apply to all people in crowds monitored by the system.[408]

The ubiquity of video recording and imaging technology, combined with the social importance placed on the visibility of faces in Western cultures, makes facial recognition a significant application of computer vision. In addition to biometric authentication, facial recognition has three other main functions: detecting faces in images or video footage (which makes it a special case of object recognition), matching faces to those recorded in a dataset, and associating an identity with a specific face.).[409] This technology raises many privacy concerns, such as the possibility of unintended uses of the information obtained by identifying who is in an image or video, how long and by whom the data is being stored by, whether the information gathered may be used in a different context, and the lack of consent or even awareness that someone's presence and activity is being recorded.[410]

An example of several of these concerns is how social media platforms such as Facebook employ facial recognition to identify people shown in photos posted by users and suggest tagging the photos with their names.[411] This may reveal information about someone's activities and location that they would prefer not to disclose to others. This example demonstrates the possibilities of unintended uses (being identified in other people's photos), a change in context due to being publicly identified in a photo on social media, and lack of awareness if the individual was unaware that a photo had been taken.

The inconspicuous addition of video cameras to many digital devices also contributes to the individual's potential lack of awareness that she has been recorded and identified, and for what purpose this information will be used for. For example, digital signs may incorporate a small video camera that records the faces of people walking past it to determine whether they looked at the sign, how long they observed it, and their likely emotional state.[412] Those appearing in such footage have no way of knowing whether facial recognition is being used to identify them or if only their response to the advertisement is being recorded.

6.1.4. Natural language processing systems

Natural Language Processing (NLP) systems are an important class of products of artificial intelligence that revolve around the processing of speech and text. They are used for applications such as analysing, translating or summarizing texts. NLP can be subdivided in Natural Language Understanding[413] (NLU)

[407] Macnish, K., 'Unblinking Eyes: The Ethics of Automating Surveillance', *Ethics and Information Technology*, Vol. 14, No. 2, pp. 151-167, June 2012.

[408] Ibid, pp. 158-159.

[409] Cammozzo, A., 'Face Recognition and Privacy Enhancing Techniques', in Bissett, A., Bynum, T. W., Light, A., Lauener, A., and Rogerson, S., ETHICOMP 2011: The Social Impact of Social Computing, Sheffield, UK, Sheffield Hallam University, pp. 101- 109, 2011.

[410] Ibid.

[411] Norval, A., and Prasopoulou, E., 'Public Faces? A Critical Exploration of the Diffusion of Face Recognition Technologies in Online Social Networks', *New Media & Society*, Vol. 19, No. 4, pp. 637-654, April 2017.

[412] Cammozzo, A., 'Face Recognition and Privacy Enhancing Techniques', in Bissett, A., Bynum, T. W., Light, A., Lauener, A., and Rogerson, S., ETHICOMP 2011: The Social Impact of Social Computing, Sheffield, UK, Sheffield Hallam University, pp. 101- 109, 2011.

[413] Although the word 'understand' implies that the algorithm 'knows' what the sentence is about, 'understand' in this context means that it tries to find a relation between words and their role in a sentence. By doing so, the algorithm is able to apply a certain weight to words, 'understanding' the importance of the role of a word in a sentence.

and Natural Language Generation (NLG).[414] Both NLU and NLG try to relate human language and "internal computer representations of information."[415] NLU does so starting from human language and relating this to the computer, and NLG starts from the internal computer information and translates this back to human language. Although they may seem to overlap, it is not easy to implement a working method working for both processes.[416] This is due to the fact that the issues arising in NLU and NLG are not necessarily the same. An important problem in NLU is that incorrect grammar should be treated as if it were correct grammar and different paraphrases should be understood to mean the same thing. For NLG, this is not an issue. In NLG a major concern regards the need for *human* understanding, something that does not arise in NLU.[417] Text Analytics or Processing is a part of NLU that tries to conceptualize the meaning of a text, including discovering "new, previously unknown information, by automatically extracting information from different written resources."[418] It is widely used in the business world, as it shows patterns unclear to human eyes, enabling "decision-makers to understand market dynamics, predict outcomes and trends, detect fraud and manage risk."[419] NLU looks at the word on its own, the role of the word in a sentence and the whole sentence in the broader perspective of the text. A major subfield of NLG is concerned with translating languages. Summarizing systems are both related to NLU as NLG. Generalized text summarization systems is difficult to build, due to the diversity of language.[420] Speech is both the most natural as the fastest way for human interaction.[421] Therefore, speech recognition systems are assumed to speed up interaction between humans and machines, as well as naturalize this interaction.[422] Voice recognition may seem similar to speech recognition, but they focus on different things. Speech recognition is focused on deciphering spoken sentences, allowing it to follow commands, answer queries, and so forth. Voice recognition on the other hand implies identifying a specific voice, thereby identifying a person.

The ethical discussion concerning natural language processing has only just started. One of the reasons given by Hovy and Spruit[423] for the lack of this discussion is that "NLP research has not directly involved human subjects", as texts used for NLP applications were usually distanced from their authors either in temporal context or in clarity about who the author was precisely.[424] Nowadays, however, there are more and more NLP applications that involve the use of social media data. This implies that both the temporal distance as the uncertainty of the author are disappearing, raising the need for an ethical discussion.[425] The remainder of this subsection first addresses the general ethical issues that are present in all of the aforementioned categories, and then touches on several issues that are specific to some of the aforementioned categories of NLP systems.

[414] Reiter, Ehud, and Robert Dale, *Building natural language generation systems,* Cambridge university press, 2000.
[415] Ibid.
[416] Ibid.
[417] Ibid., p. 3
[418] DialNet - Text Analytics, p. 1
[419] Ibid.
[420] Goldstein, Jade, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell, "Summarizing text documents: sentence selection and evaluation metrics," In *SIGIR*, Vol. 99, no. 8, pp. 121-128, 1999.
[421] Ayadi, Moataz El, Mohamed S. Kamel, and Fakhri Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases," *Pattern Recognition,* Vol. 44, No. 3, 2011, pp. 572–587., p. 572
[422] Ibid.
[423] Hovy, Dirk, and Shannon L. Spruit, "The Social Impact of Natural Language Processing," *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2016.
[424] Ibid., p. 2
[425] Ibid.

Some of the main general concerns surrounding NLP relate to privacy. Especially concerning NLP used for clinical use, data sensitivity and privacy play a big role.[426] This inhibits the progress of NLP, as access to data sets is very limited. The reason why NLP may be considered problematic for those that wish to preserve their privacy is that NLP systems are able to categorize individuals into specific groups based on the relation between language and individual traits.[427] NLP uses text bodies that contain sensitive language. As social media is increasingly used, the distance between the text and the author is reducing, making text analysis more privacy sensitive.[428] Based on the characteristics of the language used it is possible to (at least partly) identify the author. The author's living area can be conceptualized, as well as their age group or ethnicity. Research shows that anonymization of data remains the exception rather than the norm.[429] Thus, NLP tools may be used to de-anonymise people. This may be regarded as a violation of someone's privacy, especially if data is used without permission.

Speech recognition systems (e.g. 'Amazon Echo', 'Google Home') also raise privacy concerns. Such systems need to recognize when a human is speaking. To do so, usually a key phrase (e.g. 'OK, Google') activates the device. This, however, implies that the machine remains in an "always-on" mode,[430] waiting for the command to be given.[431] Such systems are not without error, however, and may start recording the user after believing to have picked up the trigger word, while in fact it was not said. The always-on mode then may be seen as a potential intrusion on someone's privacy. Furthermore, the always-on mode creates the possibility for 'hacking'. It allows "attackers to try to issue unauthorized voice commands to these devices."[432] Such systems can pick up voice commands that are unrecognizable (and therefore unnoticed) by humans.[433]

Other concerns in relation to NLP are the potential for bias and discrimination. Demographic factors commonly have been neglected in the development of NLP methods, as language was treated as a uniform phenomenon in NLP tasks.[434] The increased use of social media for developing NLP tools now shows that excluding demographic factor reduces accuracy. The main data source for NLP development is based on newswire. However, this source addresses a group that is "older, richer, and

[426] Suster, Simon, Stephan Tulkens, and Walter Daelemans, "A Short Review of Ethical Challenges in Clinical Natural Language Processing," *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 2017.
[427] Hovy, Dirk, and Shannon L. Spruit, "The Social Impact of Natural Language Processing," *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2016.
[428] Leidner, Jochen L., and Vassilis Plachouras, "Ethical by Design: Ethics Best Practices for Natural Language Processing," *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 2017., p. 6
[429] Mieskes (2017) has done a quantitative analysis on data used for NLP systems. The research involved "how often data is being collected, how data is published, and what data types are being collected" (p. 23). Reporting anonymization of data is uncommon. Mieskes' research shows how only a small percentage explicitly report the anonymization of the data. For some, it is clear that no anonymization has been done, while for others it is left in the middle. This indicates the problem of privacy. Stats: "Out of 704 publications about 32.8% collected or used data from social media or otherwise sensitive data as outlined in Section 3 above. Only about 3.5% of these report the anonymization of the data" (p. 26). See Mieskes, Margot, "A Quantitative Study of Data in the NLP Community," *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 2017.
[430] Carlini, Nicholas, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou, "Hidden voice commands," *25th {USENIX} Security Symposium ({USENIX} Security 16)*, pp. 513-530, 2016., p. 513
[431] Ibid.
[432] Ibid.
[433] Ibid., p. 525
[434] Hovy, Dirk, "Demographic Factors Improve Classification Performance," *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015., p. 752

more well-educated than the average population,"[435] thereby creating a bias. Social media, on the other hand, shows a wide diversity in age and ethnicity.[436] Misrepresentation of data may result in exclusion of the misrepresented groups.[437] Hovy and Spruit argue that this "in itself already represents an ethical problem for research purposes, threatening the universality and objectivity of scientific knowledge."[438] This bias remains a side-effect of data, and does not reside within the model itself.[439,440] (For a comprehensive discussion of biased data, see the part on "Justice and fairness" in subsection 5.1.3 of this report.)

On modelling level, bias may be maintained due to a reliance "on models that produce false positives", risking "bias confirmation and overgeneralization."[441] Additionally, research design may also lead to bias confirmation due to overexposure of a particular topic,[442] potentially leading to the concept of the "availability heuristic".[443] This implies that people appoint value to something they remember or know something about. This, however, also extends to individuals and groups. Hovy and Spruit illustrate the harm in such a heuristic when certain characteristics are linked to a specific group or ethnicity (i.e., stereotyping).[444]

Speech recognition exhibits threats of discrimination as well. Research shows racial and gender disparity in recognizing speech.[445] The accuracy of speech recognition of women and ethnic minorities is lower. For example, automatic captioning on YouTube works less well for female speakers than for male speakers[446]. This has both a negative impact on the producers of the videos as the viewers. A bias may limit the ability for speakers to share their voice with the world, as well as that other people are restricted in their information input.[447]

Further ethical issues that apply to NLP more generally relate to transparency and explainability. NLP tools are increasingly developed with neural network algorithms, thereby reducing transparency for

[435] Hovy, Dirk, and Anders Søgaard, "Tagging Performance Correlates with Author Age," *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2015., p. 483

[436] Hovy, Dirk, "Demographic Factors Improve Classification Performance," *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015., p. 752

[437] Hovy, Dirk, and Shannon L. Spruit, "The Social Impact of Natural Language Processing," *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2016., p. 593

[438] Ibid.

[439] Hovy, Dirk, and Anders Søgaard, "Tagging Performance Correlates with Author Age," *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2015., p. 487

[440] Hovy, Dirk, and Shannon L. Spruit, "The Social Impact of Natural Language Processing," *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2016., p. 593

[441] Ibid.

[442] Ibid.

[443] Ibid.

[444] Ibid., p. 594

[445] Blodgett, Su Lin, Lisa Green, and Brendan O'Connor, "Demographic Dialectal Variation in Social Media: A Case Study of African-American English," *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.

[446] Tatman, Rachael, "Gender and Dialect Bias in YouTubes Automatic Captions," *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 2017.

[447] Blodgett, Su Lin, Lisa Green, and Brendan O'Connor, "Demographic Dialectal Variation in Social Media: A Case Study of African-American English," *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016., p. 1

an improved accuracy.[448] So far, one of the best working method for NLP tools is sequence to sequence (Seq2Seq) learning, which builds forth on deep language modelling.[449] Deep language modelling uses hidden states that obscure the algorithm's visibility.

Now let us turn to some important ethical issues that are specific to different kinds of NLP systems. A first set of concerns is specific to textual analysis. NLP tools may be used to predict and nudge human behaviour.[450] This implies a potential interference with an individual's autonomy. A person may be steered towards a certain behaviour that without this nudging would not have happened. NLP tools are used to detect what factors in a text nudge a person into choosing for instance a course to study or buying a certain product. Such ideas are related to the concept of *narrative persuasion*; the influence language may have "for altering cognitive responses or attitudes."[451] This can be perceived as a reduction in someone's *autonomy* and *freedom* of choice. Simultaneously, it is related to *privacy* concerns, as nudging is related to a person's individual preferences.

It is easy to overestimate the generalization capabilities of a model. Language may be paraphrased in different ways, while still containing the same meaning. A frequent problem that follows this is called oversensitivity.[452] If a sentence in the training data has a very different structure than one the model encounters in real life, the sentence may be interpreted in a completely different way, while it actually has the same meaning. This implies that the test and training data are too similar, therefore not allowing for inclusion of "real-world" data.

Another set of concerns are specific to machine translation. Neural Machine Translation (NMT) has significantly increased the accuracy in translation. Nonetheless, it still causes problems, sometimes with severe consequences. Wrong translations cause companies a lot of money to repair the mistake. [453] While such a translation is financially problematic for a company, it may not directly involve an ethical issue. Machine translation may cause ethical issues due to ambiguity in a sentence that needs translation. Not only does this raise a problem when the sentence needs to be translated, so does it also create a problem if the ambiguity needs to be kept. By translating the sentence in a specific way, it may become unambiguous in the other language. Either, the wrong meaning of the sentence is transferred, or the sentence had two different interpretations for a meaning. Therefore, the translation contains a certain *bias*. In such a case, it may be wise to conserve the ambiguity by altering the translation.[454]

[448] Lei, Tao, Regina Barzilay, and Tommi Jaakkola, "Rationalizing Neural Predictions," *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016., p. 1. See also section 5.1.2 on Machine Learning and section 5.1.3 on Transparency & Explainability

[449] Wiseman, Sam, and Alexander M. Rush, "Sequence-to-Sequence Learning as Beam-Search Optimization," *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.

[450] Pryzant, Reid, Kelly Shen, Dan Jurafsky, and Stefan Wagner, "Deconfounded Lexicon Induction for Interpretable Social Science," *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018., p. 1.

[451] Ibid.

[452] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin, "Semantically equivalent adversarial rules for debugging nlp models," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 856-865, 2018., p. 856

[453] Zheng, Wujie, Wenyu Wang, Dian Liu, Changrong Zhang, Qinsong Zeng, Yuetang Deng, Wei Yang, Pinjia He, and Tao Xie, "Testing untestable neural machine translation: An industrial case," *Proc. 41st International Conference on Software Engineering: Companion, Poster*, 2019., p. 1

[454] Knight, Kevin, and Irene Langkilde, "Preserving ambiguities in generation via automata intersection," *AAAI/IAAI*, pp. 697-702, 2000.

Furthermore, a distinction between over- and undertranslation can be made. Undertranslation occurs when a sentence or paragraph omits words in the translated text, thereby reducing comprehensibility. Overtranslation occurs when certain words or sentences are translated more times than stated in the original text.[455]

Although there exists a general bias in NLP, the use of word embeddings poses a threat to expose existing biases in society specifically relating to NLG. Not only may certain biases in society be explicated, but some may also be reinforced. This is exemplified by translations from Turkish to English. Turkish does not use gender pronouns, and therefore "he is a doctor" is the same as "she is a doctor." In English, however, a sentence in Turkish that means "X is a doctor" is translated to "*he* is a doctor", while a sentence that reads "X is a nurse" is translated to "*she* is a nurse."[456] A common method in NLP to represent text data uses word embeddings. Word embeddings show relations between words. Technically, they are "distributed representations of words in a vector space, capturing syntactic and semantic regularities among the words."[457] Thus, these vectors relate similar meanings between words. For example: "man is to X as woman is to Y" (with x being king). An appropriate Y would then be "queen." Nonetheless, these embeddings inhibit gender biased characteristics. So is also given "man is to computer programmer as woman is to homemaker."[458]

Finally, let us turn to an issue that is specific to speech recognition systems, namely the general inability of these systems to recognise human emotions. Speech systems cannot yet comprehend emotions, which makes speech interaction between humans and machines difficult. Part of the reason is that most of the databases concerning emotional human speech are not publicly available.[459] This also has as a consequence that similar mistakes are repeated by different research groups, due to a "lack of coordination."[460]

6.1.5. Affective computing systems

Affective computing systems are systems capable of detecting, recognizing, interpreting, simulating and responding to human emotions. An application of these systems that raises concern is in biometric identification, whereby cameras and sensors become trained to go beyond simply matching individuals' faces to images in a database, to predicting 'a person's motives and/or emotional state and subsequent behaviour' by combining multi-modal input including visual, auditory, physiological and kinaesthetic variables.[461] These systems could be deployed for observing individuals in stressful situations (such as monitoring pilots) that can decide to alert supporting staff to recommend

[455] Zheng, Wujie, Wenyu Wang, Dian Liu, Changrong Zhang, Qinsong Zeng, Yuetang Deng, Wei Yang, Pinjia He, and Tao Xie, "Testing untestable neural machine translation: An industrial case," *Proc. 41st International Conference on Software Engineering: Companion, Poster*, 2019., p. 3

[456] See https://www.unleashgroup.io/news/ai-recruitment-tools-what-lies-beneath/. Note that this has now been altered in Google Translate: both he and she translations are given simultaneously.

[457] Hovy, Dirk, "Demographic Factors Improve Classification Performance," *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015., p. 754

[458] Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," In *Advances in neural information processing systems*, pp. 4349-4357, 2016.

[459] Ayadi, Moataz El, Mohamed S. Kamel, and Fakhri Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases," *Pattern Recognition,* Vol. 44, No. 3, 2011, pp. 572–587.

[460] Ibid., p. 574

[461] Bullington, Joseph, "Affective computing and emotion recognition systems: The future of biometric surveillance?," *Information Security Curriculum Development Conference '05*, 2005.

intervention given the individual's affective state, or monitor groups in collaborative situations to provide feedback on the overall emotional state of the group in an anonymized manner as part of Group Decision Support Systems.[462] While Bullington points out that at the time of writing (i.e., in 2005) these systems may not be sophisticated enough to be put in use, his aim is to bring up the potential for privacy invasion, not just of individuals in the workplace but also over the general population in urban settings. Further, the predictive power of these systems will rely on sustained tracking and learning of an individual's habits and expressions in order to be better able to infer correctly about the individual's psychological state and intentions, as they are experienced in varying situations. It will thus be necessary to consider whether this training will mean that individuals will need to self-report to the systems in order to verify the inferences these systems make, as well as ensuring that the self-reporting can be trusted so that the systems will not be gamed (i.e., tricked by those who know how the recognition and classification works).[463]

The pervasiveness of these systems therefore raises issues of privacy and trust. If these systems become capable of correctly inferring the emotional states of individuals and groups, as well as displaying emotional states of their own, one potential situation is that social bonds become forged between humans and these affective systems.[464] On the one hand, these human-machine bonds may be useful in contexts where individuals may be suffering from loneliness and isolation (such as is the case with elderly individuals in care home situations).[465] But on the other hand, the better such systems are at persuading and making individuals believe that they are actually exhibiting human-like emotions and affects (such as with the Tamagotchi, or Paro the robot seal which arouse an emotional connection), the more likely they may potentially be emotionally manipulative by purposefully modifying the human user's emotional state.[466] And such human-machine intimacy is only possible if affective systems are able to be part of an individual's personal life, meaning that privacy and trust in these systems needs to be addressed as well when considering how and where these systems are deployed. As these systems are capable of retrieving, storing and disseminating emotional data from the users they are monitoring and tracking, new forms of privacy protection will be needed to ensure this information is not used against the individuals who are under the gaze of these systems.[467,468]

An additional concern stemming from how convincingly emotive affective computing systems may be is that human users may end up being deceived by these systems. Deception can take place when affective systems persuade or prompt their human users to make decisions that may not be in their best interest.[469] The more likely these systems can evoke emotional competence, but do not have the necessary conscience to be aware of whether or not certain actions should or should not be followed to prompt human users, the easier it will be for these systems to successfully persuade users.[470] Systems that can effectively mimic emotional responses and that lead individuals to believe that the

[462] Ibid.
[463] Ibid.
[464] Duffy, Brian, "Fundamental Issues in Affective Intelligent Social Machines," *The Open Artificial Intelligence Journal*, No. 2, 2004, pp. 21–34.
[465] Ibid.
[466] Ibid.
[467] Ibid.
[468] Reynolds, Carson, and Picard, Rosalind, "Affective Sensors, Privacy, and Ethical Contracts," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI 2004 Extended Abstracts on Human Factors in Computing Systems, pp. 1103–1106, Vienna, Austria, ACM, 2004.
[469] Cowie, Roddie, "Ethical issues in affective computing," In *The Oxford handbook of affective computing*, Oxford Library of Psychology, 2015.
[470] Ibid.

systems are in their best interest, can thus not only carry the risk of being emotionally manipulative but also the risk of potentially deceiving individuals into making financial decisions that may not be in their interest due to how persuasive they may end up becoming. Another application context where this issue may arise is in cases of artificial agents that are deployed as teachers or companions, built to exhibit affective states so they may mimic the capacity of caring. If users assume that these agents will do more than what they are specifically built for (to issue guidance on courses or to be conversational), they may end up being frustrated and feel deceived when they realise that these systems cannot do more, an issue that might arise in children or individuals with cognitive impairments.[471] In such situations, the level of intelligence of these systems (regarding how much the decisions they prompt users to take are scripted or made from how the system learns through experience) will reflect where responsibility for such deception shall fall - either on the systems, or on their designers.[472]

Coupled to the risks of emotional manipulation and deception is the potential for these systems to diminish the autonomy of individuals. The fact that affective computing systems can store and act upon emotion data from human users, as well as portray affective states themselves, means that it is possible for these systems to affect how individuals reach their decisions. The very aim of affective computing, to affect people's feelings, entails the danger that once individuals have their emotions influenced then their decision-making and self-reflective capacities will be compromised.[473] As such, the more affectively competent such systems are, and the more likely individuals are to form social bonds that lead to trusting these systems, the greater the risks of autonomy being endangered because it will be easier to accept the prompts of these systems than be critical of them.

6.1.6. (Big) Data analytics systems

One of the most researched, utilized and invested in technologies in present computing infrastructure is Big Data. While the term may imply merely that the size of data at the disposal of businesses, governments and individuals has become remarkable, the importance of this technology lies in the capacities that Big Data systems offer beyond just the volume of data stored and used. There are a number of key structural traits that distinguish Big Data: (1) high in volume, (2) high in velocity with data being created in or near real-time, (3) exhaustive in scope, aiming to capture information about entire systems or populations, (4) indexical in identification, being able to store more information about each piece of data, (5) relational in nature to allow connections between data sets, (6) flexible, being able to add new data easily and datasets can be expanded in size rapidly.[474] These capacities are made possible by incremental developments in data storage, data mining techniques, ubiquitous computing infrastructure and connectivity of computing devices. These capacities of Big Data allow more data from individuals and groups to be stored, as well as the development of algorithms (in the domain of Big Data analytics) to make sense of the data for inferring patterns to inform decisions for businesses, governments as well as individuals and groups. It is because of the increased implementation of Big Data analytics and Big Data systems in economic, educational, medical, personal and governance related decision making that ethical attention is necessary.

One of the key areas of concern with the ascent of Big Data is how these systems problematise privacy concerns. One the one hand, the more information that Big Data analytics and Big Data systems have access to, the more accurate decisions reached may become (for instance recommendation algorithms

[471] Ibid.
[472] Ibid.
[473] Ibid.
[474] Robert Kitchin, (2014). 'Big Data, new epistemologies and paradigm shifts' in *Big Data & Society*, 1(1): 1-12.

and personalised advertisements) and the more value can be generated for these systems.[475] But on the other hand, as data from individuals and groups is captured and stored by public and private institutions, this points towards an additional issue, namely the potential for increased surveillance of individual and group behaviour and associated privacy concerns. There is thus a need to evaluate how to properly utilize Big Data, either for economic benefit, democratic ideals, or for surveillance purposes.[476] This is so especially as Big Data represents the increased pervasiveness of data collection and storage techniques, tracking of individual and group behaviour from multiple sources (e.g., smartphones, wearables, social media),[477] and the constant connectivity borne from ICT infrastructures.[478] This therefore has implications for defining how users should be informed about the capture, storage and use of their data. Furthermore, infringement of individual and group privacy may also be considered as infringements on human dignity and autonomy,[479] which makes the potential for surveillance all the more significant as an ethical concern.

The invasion of individual and group privacy through surveillance and tracking of their behaviour may lead to concerns over responsibility and accountability. Reaching a consensus on who is responsible when an algorithmic decision leads to negative consequences is problematic for a number of reasons, especially in the case of Big Data systems. Due to the complexity of these systems, the technical opacity of algorithms (i.e. difficulty in interpreting the decision-making process), and algorithms being considered as decision makers, there is an "accountability gap" between human designers, algorithms and institutions.[480] The opacity creating this accountability gap can stem from corporate self-protection, technical illiteracy, and the contrast between what the algorithm looks like and what our practical reasoning expects as explanation for a decision.[481] These three forms of opacity make it more difficult for regulatory bodies as well as researchers to identify biases in the design and training of algorithms, which mean identifying issues is usually only after the decisions have been made.

Another notable area of concern in the deployment of Big Data analytics is the potential for discriminatory decision-making. Such discrimination can arise when the decisions that are made can

[475] Manuel Souto-Otero and Benito-Montagut, R. (2016). 'From governing through data to governmentality through data: Artefacts, strategies and the digital turn' in *European Educational Research Journal* Vol. 15(1): 14-33; Luke Hutton & Henderson, T. (2017). 'Beyond the EULA: Improving Consent for Data Mining' in *Transparent Data Mining for Big and Small Data*, (Ed.) Tania Cerquitelli, Daniele Quercia and Frank Pasquale. Springer.
[476] Sami Coll. (2014). Power, knowledge, and the subjects of privacy: understanding privacy as the ally of surveillance. *Information, Communication & Society*, 17(10), 1250-1263; Gordon Hull, (2015). Successful failure: what Foucault can teach us about privacy self-management in a world of Facebook and big data. Ethics and Information Technology, 17(2), 89-101; Omar Tene and Polonetsky, J. (2012). Big data for all: Privacy and user control in the age of analytics. *Northwestern Journal of Technology & Intellectual Property*, 11(5): 238-273.
[477] Alexander R. Bentley, O'Brien, M., J. & Brock, W. A., (2014). Mapping collective behavior in the big-data era. Behavioral and Brain Sciences 37 (1):63-76.
[478] Jenna Burrell, (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1).
[479] Manon Oostveen and Irion, K. (2018). 'The Golden Age of Personal Data: How to Regulate an Enabling Fundamental Right?' in M. Bakhoum, B. Conde Gallego, M-O. Mackenrodt, & G. Surblytė-Namavičienė (Eds.), *Personal Data in Competition, Consumer Protection and Intellectual Property Law: Towards a Holistic Approach?* (pp. 7-26). (MPI Studies on Intellectual Property and Competition Law; Vol. 28). Berlin: Springer.
[480] Burrell, op. cit., 2016; Beatriz Cardona (2008) 'Healthy ageing' policies and anti-ageing ideologies and practices: On the exercise of responsibility. Medicine, Health Care and Philosophy 11(4): 475–483; Andreas Matthias (2004) The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology 6(3): 175–183; and Tal Zarsky (2016) The trouble with algorithmic decisions an analytic road map to examine efficiency and fairness in automated and opaque decision making. Science, Technology & Human Values 41(1): 118–132.
[481] Burrel, op. cit., 2016 p1-2.

be shown to be biased against a particular individual or group based on ethnicity, gender or belonging to that particular group[482] such as in cases of predictive policing,[483] employee hiring[484] and credit scoring. This discrimination may occur either consciously or subconsciously in the design of the algorithms that are used to make decisions from analysis of the data at hand. Friedman & Nissenbaum (1996) outline three types of algorithmic biases that may lead to discriminatory decision making, these are: pre-existing bias (biases in society perpetuated in algorithms), technical bias (false correlations that lead to negative effects from decisions), and emergent bias (arising when a system is in use). Discrimination is therefore an issue that emerges from the design, training and use of Big Data analytics and Big Data systems. And the multiple domains that Big Data is utilized in, mean that discrimination can lead to negative impacts on justice, fairness, equality and autonomy, as decisions made by algorithms can have adverse effects on how individuals and groups are treated based on how they are classified in these systems. (Big data is affected by the same biases as AI and computer systems generally. See the part on "Justice and fairness" in subsection 5.1.3 for a more substantial treatment of algorithmic bias.)

6.1.7. Embedded AI and Internet of Things

The concept of Internet of Things (IoT) refers to the interconnection via the Internet of computing devices (which are often embedded in everyday objects) that enables these devices to share and exchange data without requiring human-to-human or human-to-computer interaction. IoT is generally unable to fulfil its promises by itself as it is unable to make sense of the data that is communicated. For this, it often requires artificial intelligence, specifically machine learning algorithms. Such algorithms can interpret the data collected by the IoT to provide a deeper understanding of hidden patterns within the data. This allows the devices to for instance adapt to users' preferences or provide predictive maintenance (i.e., prevent possible harms by analysing patterns). Devices that combine IoT and AI are also referred to as "smart devices". Smart devices aim to assist people in their life using technologies embedded in the environment. Some important characteristics of such a device include that it is embedded (device is "invisible" to the user), context-aware (device recognizes users), personalized (device is tailored to user's need), adaptive (device is able to change according to its environment and/or user), anticipatory (device can anticipate a user's desires), unobtrusive (device is discrete) and non-invasive (device can act on its own, does not necessarily require user's assistance).[485]

Related to the Internet of Things are embedded systems. An embedded system commonly requires little to no human interference and provides a connection between devices. The Internet of Things is a specific type of an embedded system, namely in which the devices are connected through the internet. Applying artificial intelligence to embedded systems creates the concept of *Embedded AI.* Embedded AI is not limited to embedded systems that are connected through the internet.

[482] Toon Calders, Kamiran, F and Pechenizkiy, M (2009) Building classifiers with independency constraints. In: Data mining workshops, 2009. ICDMW'09; Cohen IG, Amarasingham R, Shah A, et al. (2014) The legal and ethical concerns that arise from using complex predictive analytics in health care. Health Affairs 33(7): 1139–1147; and Anthony Danna and Gandy OH Jr (2002) All that glitters is not gold: Digging beneath the surface of data mining. Journal of Business Ethics 40(4): 373–38.

[483] Joni R. Jackson, (2018). Algorithmic Bias. Journal of Leadership, Accountability and Ethics, 15(4), 55-65.

[484] Amit Datta, Tschantz M. C. & Datta, A. (2015). Automated Experiments on Ad Privacy Settings A Tale of Opacity, Choice, and Discrimination. Proceedings on Privacy Enhancing Technologies 2015 (1), 92–112.

[485] Gams, Matjaz, Irene Yu-Hua Gu, Aki Härmä, Andrés Muñoz, and Vincent Tam, "Artificial Intelligence and Ambient Intelligence," *Journal of Ambient Intelligence and Smart Environments,* Vol. 11, No. 1, 2019, pp. 71-86., p. 76.

Devices that combine embedded AI with IoT have the following characteristics: ubiquitous computing, ubiquitous communication and user adaptive interface.[486] Ubiquitous computing, a term coined by Mark Weiser, refers to "computer use by making computers available throughout the physical environment, while making them effectively invisible to the user."[487] It aims to "serve people in their everyday lives at home and at work, functioning invisibly and unobtrusively in the background and freeing people to a large extent from tedious routine tasks."[488] Ubiquitous communication implies that computers have the ability to interact with each other. This can also be seen as a part of ubiquitous computing.

A user adaptive interface, or *intelligent social user interface* (ISUI) has as its main characteristics profiling ("ability to personalize and automatically adapt to particular user behaviour patterns") and context-awareness ("ability to adapt to different situations"[489]). Devices with the ISUI component are able to "infer how your behaviour relates to your desires."[490] ISUI includes the ability to recognize visual, sound, scent and tactile outputs.[491]

IoT and Embedded AI have several benefits, such as the potential to save people time and money, provide a more convenient life, and increase the level of safety, security and entertainment.[492] This, then, may lead to "an overall higher quality of life."[493] Although some, if not all, of these benefits are likely, several ethical concerns arise with their usage, relating to privacy, identity, trust, security, freedom and autonomy.[494] Furthermore, smart technologies may influence people's individual behaviour as well as their relation to the world.[495,496]

Privacy concerns are considered of utmost importance by both critics and proponents of embedded AI and IoT technologies.[497] Four properties of ubiquitous computing that make it especially privacy sensitive compared to other computer science domains include ubiquity, invisibility, sensing, and memory amplification.[498] Thus, ubiquitous computing is everywhere, unnoticed by humans, with the ability to sense aspects of the environment (e.g. temperature, audio) as well as of humans (e.g.

[486] Raisinghani, Mahesh S., Ally Benoit, Jianchun Ding, Maria Gomez, Kanak Gupta, Victor Gusila, Daniel Power and Oliver Schmedding, "Ambient intelligence: Changing forms of human-computer interaction and their social implications," *Journal of digital information,* Vol. 5, No. 4, 2004.

[487] Weiser as cited in Spiekermann, Sarah, and Frank Pallas, "Technology paternalism–wider implications of ubiquitous computing," *Poiesis & praxis,* Vol. 4, No. 1, 2006, pp. 6-18., p. 7

[488] Ibid.

[489] Soraker, Johnny Hartz, and Philip Brey, "Ambient intelligence and problems with inferring desires from behaviour," *International Review of Information Ethics,* Vol. 8, no. 1, 2007, pp. 7-12., p. 8

[490] Ibid., p. 9

[491] Raisinghani, Mahesh S., Ally Benoit, Jianchun Ding, Maria Gomez, Kanak Gupta, Victor Gusila, Daniel Power and Oliver Schmedding, "Ambient intelligence: Changing forms of human-computer interaction and their social implications," *Journal of digital information,* Vol. 5, No. 4, 2004.

[492] Ibid.

[493] Ibid.

[494] Wright, David, "The Dark Side of Ambient Intelligence," *Info,* Vol. 7, No. 6, 2005, pp. 33–51., p. 34; Brey, Philip, "Freedom and Privacy in Ambient Intelligence," *Ethics and Information Technology*Vol. 7, No. 3, 2005, pp. 157–166., p. 4

[495] Soraker, Johnny Hartz, and Philip Brey, "Ambient intelligence and problems with inferring desires from behaviour," *International Review of Information Ethics,* Vol. 8, no. 1, 2007, pp. 7-12.

[496] Araya, Agustin A., "Questioning Ubiquitous Computing," *Proceedings of the 1995 ACM 23rd Annual Conference on Computer Science - CSC 95*, 1995., p. 236

[497] Brey, Philip, "Freedom and Privacy in Ambient Intelligence," *Ethics and Information Technology,* Vol. 7, No. 3, 2005, pp. 157–166., p. 8.

[498] Langheinrich, Marc, "Privacy by Design — Principles of Privacy-Aware Ubiquitous Systems," *Ubicomp 2001: Ubiquitous Computing Lecture Notes in Computer Science*, 2001, pp. 273–291., p. 6.

emotions) and potentially creating "a complete record of someone's past."[499] Regarding the Social Interface, one may add the properties of profiling (i.e. constructing unique profiles of users) and connectedness (wireless connection between devices).[500] The privacy risks of embedded AI and IoT are considerable due to the aspect of interaction between devices. It is the combination of the sensitivity of the recorded information, the scale of this recording, and the possibility that interaction of devices facilitates distribution of personal information to other parties that makes embedded AI and IoT so vulnerable to privacy violation.[501] Relating to privacy concerns are concerns about security of, and trust in, embedded AI and IoT systems. Trust is important for all human-technology relations.[502] If a user has the feeling that the system may have malicious intentions, he or she might be reluctant to use the system. It is thus essential that the user can trust the system.

While IoT and embedded AI may be regarded as fostering freedom due to time and money savings, it may also be regarded as diminishing human autonomy and freedom.[503] Autonomy is commonly regarded as dependent on an individual's ability to make their own decisions, and is seen as important due to the opportunity for "self-realisation."[504] Furthermore, freedom and autonomy are closely related. Freedom may be split in two categories; no one must stand in your way, and no one should tell you what to think.[505] Brey (2005) has analysed the concept of IoT and AI in relation to these types of freedoms, and concludes that IoT combined with AI has a chance to enhance our freedom in both ways: it may "enhance control over the environment by making it more responsive to one's needs and intentions" as well as improve "our self-understanding and thereby helping us become more autonomous."[506] It simultaneously limits both freedoms by confronting "humans with smart objects that perform autonomous actions against their wishes and "by pretending to know what our needs are and telling us what to believe and decide."[507]

In addition, the use of IoT and embedded AI systems may influence a person's behaviour.[508] Søraker and Brey argue that for IoT and embedded AI systems to understand what we want, the behaviour humans need to show to a device is similar to the behaviour they need to show to a pet; it must be "discrete, predictable and overt."[509] They claim that this may change our natural behaviour. Thus, IoT and embedded AI may force us into changing who we are and how we act; we will then be forced to fit ourselves within this technology. Moreover, some IoT and embedded AI devices may promote their use in solitude, risking isolation of individuals and a degeneration of society. Also, as some devices may replace tasks as doing groceries, the "face-to-face interaction between people" might diminish,[510] potentially adding to a feeling of isolation. Furthermore, as IoT and embedded AI technologies spread

[499] Brey, Philip, "Freedom and Privacy in Ambient Intelligence," *Ethics and Information Technology,* Vol. 7, No. 3, 2005, pp. 157–166., p. 9.
[500] Ibid.
[501] Ibid.
[502] Tzafestas, Spyros, "Ethics and law in the internet of things world," *Smart Cities*, Vol. 1, no. 1, 2018, pp. 98-120., pp. 112-115
[503] Brey 2005, p. 4.
[504] Ibid.
[505] Ibid.
[506] Ibid., p. 8.
[507] Ibid.
[508] Soraker, Johnny Hartz, and Philip Brey, "Ambient intelligence and problems with inferring desires from behaviour," *International Review of Information Ethics,* Vol. 8, no. 1, 2007, pp. 7-12.
[509] Ibid., p. 10.
[510] Raisinghani, Mahesh S., Ally Benoit, Jianchun Ding, Maria Gomez, Kanak Gupta, Victor Gusila, Daniel Power and Oliver Schmedding, "Ambient intelligence: Changing forms of human-computer interaction and their social implications," *Journal of digital information,* Vol. 5, No. 4, 2004.

globally, there is a risk of cultural bias. This may result in discrimination of some cultures and encourage "homogenization of cultural expressions."[511] Finally, IoT and embedded AI systems may lack easy to access and easy to use manual overrides. Søraker and Brey warn for a potential widening between users that simply go along with the requirements of the device and people that try to "game the system."[512] Not only is there an influence on individual level, it has been argued that the whole relation between men and world may be altered, as the entire world is transformed into a surveillable object.[513]

Finally, some other concerns relate to responsibility and accountability. Who decides what the device shares and records? [514] Perhaps the device acts in a way unintended by the designer and unwanted by the user. Who is to blame in such a case?[515]

6.2. Ethical issues with robotics products

This subsection identifies and describes the potential ethical issues that are either inherent in, or may occur across a wide range of applications of, important kinds of robotics products. It discusses, in turn, the issues for *humanoid robots* (subsection 6.2.1), *social robots* (subsection 6.2.2), *unmanned aerial vehicles* (subsection 6.2.3), *self-driving vehicles* (subsection 6.2.4), *telerobotic systems* (subsection 6.2.5), *robotic exoskeletons* (subsection 6.2.6), *biohybrid robots* (subsection 6.2.7), *swarm robots* (subsection 6.2.8), *microrobots* (subsection 6.2.9), and *collaborative robots* (subsection 6.2.10). Table 9 below lists the most important ethical issues that have been identified for each of these types of robotics products.

Type of product	Ethical issues	
Humanoid robots	**- Misplaced moral accountability** **- Misplaced trust** **- Misplaced empathy**	**- Reinforcement of stereotypes** **- Acceptance of abusive behaviour**
Social robots	**- Misplaced moral accountability** **- Misplaced trust** **- Misplaced empathy** **- Diminished social interaction** **- Reinforcement of stereotypes**	**- Acceptance of abusive behaviour** **- Bias** **- Privacy** **- Exacerbation of social inequality**
Unmanned Aerial vehicles	**- Privacy** **- Safety** **- Security**	**- Transparency** **- Responsibility and accountability** **- Permissibility in various contexts**

[511] Soraker, Johnny Hartz, and Philip Brey, "Ambient intelligence and problems with inferring desires from behaviour," *International Review of Information Ethics,* Vol. 8, no. 1, 2007, pp. 7-12., p. 11.

[512] Gaming the systems entails that someone may understand how a device responds to a user's behaviour, and therefore intentionally behaves in a specific way to conform the device to his/her own desires. This is problematic if a device is not merely for individual use, but rather for an embedded AI device meant to be used by for multiple people. See Soraker & Brey, 2007, p. 11.

[513] Araya, Agustin A., "Questioning Ubiquitous Computing," *Proceedings of the 1995 ACM 23rd Annual Conference on Computer Science - CSC 95*, 1995., p. 235.

[514] Bohn, J., V. Coroamă, M. Langheinrich, F. Mattern, and M. Rohs, "Social, Economic, and Ethical Implications of Ambient Intelligence and Ubiquitous Computing," *Ambient Intelligence*, 2005, pp. 5–29.

[515] Matthias, Andreas, "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata," *Ethics and Information Technology,* Vol. 6, No. 3, 2004, pp. 175–183.

Self-driving vehicles	- **Privacy** - **Safety** - **Security**	- **Transparency** - **Responsibility and accountability** - **Design of crash algorithms**
Telerobotic systems	- **Diminished social interaction** - **Psychological well-being** - **Increased "technologisation"** - **Safety**	- **Security** - **Exacerbation of social inequality** - **Responsibility and accountability**
Robotic exoskeletons	- **Safety** - **Physical wellbeing** - **Psychological wellbeing** - **Access and equality**	- **Privacy** - **Security** - **Dehumanisation**
Biohybrid robots	- **Moral status and permissibility**	
Swarm robots	- **Privacy** - **Safety**	- **Security** - **Potential for military applications**
Microrobots	- **Privacy** - **Control and ownership**	- **Safety** - **Environmental degradation**
Collaborative robots	- **Trust** - **Psychological wellbeing**	- **Privacy** - **Security**

Table 9: Overview of ethical issues with major types of robotics products.

6.2.1. Humanoid robots

Humanoid robots refer to a category of robots designed to imitate human beings in appearance, mannerisms, language, emotions, and/or actions. While not limited to these areas of emulation, the primary purpose of humanoid robots is to cross the "uncanny valley"[516] and encourage humans to interact with robots as though they were interacting with another human being. As such, one of the key ethical problems that bind many objections together in humanoid robot dialogues is that of *anthropomorphism*—attributing behaviours, emotions, thoughts, or characteristics of humans to an entity incapable of possessing them. This category mistake can be surmised by attributing human mind/emotion/intentionality to something decidedly not human. The anthropomorphizing of machines can very well lead to several ethical dilemmas, namely: misplaced moral accountability, misplace trust/emotional accountability, and misplaced empathy.[517,518]

Misplaced moral accountability occurs when a human finds the robot itself morally blameworthy for an action or response. When humans mistakenly view robots as synthetic agents or patients, but do not fully grasp that its actions and responses are not entirely of its own free choice or creation. People mistakenly attribute certain responses to be the robot's "personality" or "attitude", i.e., calling Apple's Siri "sassy", but do not quite grasp that the robot has no choice in the matter. As such, it is not the fault of Siri for using banter to ward off user's sexual advances, it is just what the robot is programmed to do and "be like"—much to women's rights activists' chagrin.[519] Moral accountability for robots is

[516] This term is defined further on this subsection.
[517] Chatila, Raja, "Inclusion of Humanoid Robots in Human Society: Ethical Issues", *Humanoid Robots: A Reference*, October 2017.
[518] Polgar, David Ryan, "Is it Unethical to Design Robots to Resemble Humans?", *Quartz*, June 2017.
[519] Fessler, Leah, "We Tested Bots Like Siri and Alexa to See Who Would Stand Up to Sexual Harassment", *Quartz*, February 2017.

better directed at various points in the product's design chain (company approval, product design, product programming, etc.) than at the robot itself, as the options available, final selection, and solution execution is not consciously generated by the robot. Misplaced moral accountability is problematic as it pushes the fault onto the robot itself while turning a relatively blind eye to the rest of the robot's life cycle. This may allow robot manufacturers to skirt accountability.[520,521]

Misplaced trust/emotional accountability occurs when humans are encouraged to interact with a robot *as if* it was another human being in settings or contexts in which there is a high degree of emotional or physical intimacy—giving a robot the same access to private life as one would do with another person. This is problematic because it is not capable of imitating humans perfectly, which may cause unintended psychological consequences (to be discussed) and because robots can be capable of recording, storing, sending, and receiving data unlike a human being. As such, especially in areas of sex, companionship, or therapy, vulnerable individuals may be inclined to trust a robot and disclose heavily sensitive information. Thus, ethical questions of *privacy* (e.g., How is data being stored? When is data being collected? What type of data is being collected?), *transparency* (e.g., Do users have access to their data? Can this data be deleted? Can one "opt-out"?), *control* (e.g., Who controls this data? Who can view it? Who has rights to it? Can it be sold?), and *security* (e.g., Is the robot vulnerable to hacking?) surface that normally would not with human interactors, since individuals are generally better able to judge how trustworthy another human is to them and directly choose the degree of intimacy they wish to expose. Thus, creating machines to appear more humanlike may encourage users to divulge more sensitive data than they would normally broadcast a computer or non-anthropomorphic machine.[522,523,524]

Anthropomorphizing robots leads to an increase in empathic responses towards robots.[525] And while empathizing with robots is not bad in itself, and in fact could lead to novel insights on human interactions, abuse, and relationships, it does open the door to allowing the gaming or "hacking" of human emotions. Wherein, companies will make certain design choices for the construction of humanoid robots that elicit these empathy responses so that humans will trust the robot and grant insight into heavily intimate areas of their life. This would then lead to many of the ethical problems as outlined above.[526] In the wrong hands, these robots could glean data from individuals that would normally be inaccessible to the realms of marketing and analytics to be used in ways to current social media insights.

In the pursuit of designing humanoid robots, researchers often hit the impasse of the uncanny valley—a feeling of disconnect, 'creepiness', or mistrust that occurs in the human brain when the robot appears to be a human interactor, but the robot fails to live up to the mental predictive 'schema'

[520] Leggett, Theo, "Who is to Blame for 'Self-Driving Car' Deaths?", BBC, May 2018.
[521] Kahn Jr. Peter H., Kanda, Takayuki, & Ishiguro, Hiroshi et al., "Do People Hold a Humanoid Robot Morally Accountable for the Harm it Causes?", *Attitudes and Responses to Social Robots*, March 2012.
[522] Chatila, 2017, op. cit.
[523] Pisch, Anita, "The Ethics of Human Robots: Sam Jinks Brings an Artist's Perspective to the Discourse", *The Conversation*, October 2017.
[524] Kahn Jr., Peter H., Kanda, Takayuki, & Ishiguro, Hiroshi, et al., "'Robovie, You'll Have to Go into the Closet Now': Children's Social and Moral Relationships with a Humanoid Robot", *Developmental Psychology* 48(2), 2012.
[525] Riek, Laurel D., Rabinowitch, Tal-Chen, Chakrabarti, Bhismadev, & Robinson, Peter, "How Anthropomoprhism Affects Empathy Towards Robots", Cambridge, 2009.
[526] Chatila, 2017, op. cit.

established on what human agents ought to be like.[527] This phenomenon typically prevents even highly human-appearing robots from passing as humans in most contexts. Problematically, when it comes to vulnerable groups (children, elderly, those with cognitive disabilities), this ability to feel the effects of the uncanny valley recedes. These vulnerable groups seem less likely to experience the psychological cues that the humanoid robot is indeed a robot and does not have emotions, thoughts, desires, etc.[528,529] This is particularly concerning when social robots are being used and proposed for companionship, care, and therapy of some of these vulnerable groups.

Finally, there is the ethical concern of abuse. Not only in the case of corporate abuse, like gamifying human emotions,[530] but also in the case of acting abusively towards humanoid robots and designing robots to perpetuate abuse, like rape robots,[531,532] or designing robotic personal assistants to tolerate, accept, or avoid sexual harassment.[533] Researchers exploring these topics find cause for concern in the mistreatment of robots that elicit human responses, as the long-term impacts on human-to-human interactions are widely unstudied and may be undesirable. Further, allowing certain stereotypes to go unchecked in the designs of robots may perpetuate human stereotyping and harm by reinforcing already harmful social norms.

6.2.2. Social robots

Social robots, while often used in contexts similar to those of humanoid robots, do not necessarily seek to maintain an exceedingly human-like appearance. In some cases, social robots appear animal-like, as in the case of Boston Dynamic's "SpotMini" or AIST's PARO, or just may appear distinctly not-human, but share similar features (emotional displays with facial expressions) like Blue Frog's "Buddy" or Cognitoy's "Miko". Given the similarity of use-contexts between social robots and humanoid robots, many of the same ethical dilemmas are important—especially in contexts where robots are given positions of trust or working with vulnerable populations.[534] However, some of the risks may be lowered as social robots are not expressing human emotions with nearly the levels of authenticity as humanoid robots. As such, it would seem likely the potential for psychological gaming are decreased,

[527] Pinar Saygin, Ayse, Chaminade, Thierry & Ishiguro, Hiroshi et al., "The Thing That Should Not Be: Predictive Coding and the Uncanny Valley in Perceiving Human and Humanoid Robot Actions", *Social Cognitive and Affective Neuroscience* 7(4), April 2012.

[528] Kahn, et al., 2012, op. cit.

[529] Shamsuddin, Syamimi, Yussof, Hanafiah & Ismail, Luthffi, et al., "Initial Response of Autistic Children in Human-Robot Interaction Therapy with Humanoid Robot NAO", *IEEE*, March 2012.

[530] Polgar, 2017, op. cit.

[531] Sparrow, Robert, "Robots, Rape, and Representation", *International of Journal of Social Robotics*, May 2017.

[532] Danaher, John, "Robotic Rape and Robotic Child Sexual Abuse: Should They be Criminalised?", *Criminal Law and Philosophy*, December 2014.

[533] Fessler, 2017, op. cit.

[534] Meghdari, Ali & Alemi, Minoo, "Recent Advances in Social & Cognitive Robotics and Imminent Ethical Challenges", *Advances in Social Science, Education, and Humanities Research* 211, 2018.

while still retaining the psychological benefits of companionship, reduced stress, and mood boosts.[535,536,537]

Concerns of note are bias and inequality—both in accessibility to these robots and also in ensuring robots are not perpetuating or inventing stereotypes that negatively impact human beings.[538,539,540] Further worth repeating from the above section on humanoid robots is the need for chains of responsibility for when robots do err, regulations on how robots ought to be treated (e.g. Can a robot be prohibited from doing its job by a human without reprimand?), and further discussion on transparency and confidentiality of the data robots collect and use. In addition, the discussion of questions concerning appropriate, or inappropriate, contexts and uses for social robots would also be valuable. Should these robots be able to be substitutes for human interactions in schools, healthcare, and home life? The risk of social isolation would be an important concern if many of an individual's interactions are human-to-robot. In addition, to what extent can, and should, robots be trusted when dealing with vulnerable populations? And, lastly, are robots to be considered slaves in terms of rights and consideration (and what impact does this have on how human beings see one another)?[541]

6.2.3. Unmanned aerial vehicles (UAVs)

UAVs, also commonly known as "Drones" or Unmanned Aerial Systems (UAS), refer to a class of unmanned flying vehicles that either operate independently or as a surrogate for human controllers remotely operating them from afar. Many of these ethical issues will overlap with other unmanned or "autonomous" vehicles, but, importantly, the ethical issues at hand will very, often significantly, depending on the design, purpose, and ownership.[542] One of the key areas of concern that affects nearly all AUVs is that of acceptable use and usage locations. Problems in this area are primarily concerns of matters of authority (flying drones on the freeway, in a subway, over private property), collision (persons, other aircraft, wildlife), and ownership (persons, private corporations, institutions,

[535] Moyle, Wendy, Bramble, Marguerite, Jones, Cindy & Murfield, Jenny, "'She Had a Smile on Her Face as Wide as the Great Australian Bite': A Qualitative Examination of Family Perceptions of a Therapeutic Robot and a Plush Toy", *The Gerontologist* 00(00), October 2017.

[536] Moyle, Wendy, Bramble, Marguerite, Jones, Cindy & Murfield, Jenny, "Care Staff Perceptions of a Social Robot Called Paro and a Look-Alike Plush Toy: A Descriptive Qualitative Approach", *Aging & Mental Health* 22(3), November 2016.

[537] Arnold, Thomas & Scheutz, Matthais, "The Tactile Ethics of Soft Robotics: Designing Wisely for Human-Robot Interaction", *Soft Robotics* 4(2), 2017.

[538] Moyle, et al., 2016, op. cit.

[539] Fessler, 2017, op. cit.

[540] Howard, Ayanna & Borenstein, Jason, "The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity", *Science and Engineering Ethics* 24(5), October 2018.

[541] Meghdari, et al., 2018, op. cit.

[542] Wilson, Richard L., "Ethical Issues with Use of Drone Aircraft", IEEE International Symposium on Ethics in Science, Technology and Engineering, May 2014.

military). A police UAV will have quite different jurisdiction and collision rules than a UAV manned by a corporation or individual.[543,544,545,546,547]

One of the other notable areas for ethical consideration is that of UAV design. For example, should different UAVs have range and battery allowance dependent on use? Should there be sound regulations for commercial or personal use UAVs? Or do police, government, and corporate UAVs need to be specially marked as their various manned transportation vehicles are? Also considering other add-ons like camera strength, infrared, camouflage, and lights, which all may be unwelcome or potentially dangerous for personal use UAVs. Many of these questions will tie into the above ones on desirable outcomes for maintaining social order.[548] Another area of concern is that of pollution and wildlife interference. In some cases, these UAVs may be used to monitor environmental conditions, but if the noise or presence of the drone causes stress to the creatures around it, it may be doing more harm than permissible. Further, if drones crash and go unnoticed, they may cause wildfires, in some areas, or otherwise contribute to the already-problematic issue of e-waste.[549,550]

In relation to this, the problem of accountability and responsibility is another popular topic in UAV debates. If an individual is injured by a UAV, a fire caused by a UAV, or a person would like to file a complaint about a UAV, who is responsible for fielding these problems and fixing them?[551,552,553,554] Further, if tied into another key ethical concern of privacy and surveillance, what are the limitations of what can be recorded by a drone? No matter who is using the drone (corporations, law enforcement, individual), what can and cannot be recorded (especially without consent or a warrant) seems to be an incredibly grey area. Is recording someone's empty house a problem? Is it of moral concern to observe people from so far away one could not identify them? Are companies allowed to collect data by drone like Google Maps cars do? UAVs may end up altering privacy expectations in ways other vehicles do not—personal airspace questions, noise violations, and recording disputes abound. While many individuals may not have these expectations in public, having hordes of drones flying around one's personal home may cross a few lines, especially those with video or audio recording or lights.[555,556,557]

[543] Ibid.
[544] Dempsey, Caitlin, "Drones and GIS: A Look at the Legal and Ethical Issues", *GIS Lounge*, September 2015.
[545] Hodgson, Jarrod & Lian Pin Koh, "Best Practice for Minimising Unmanned Aerial Vehicle Disturbance to Wildlife in Biological Field Research", *Current Biology* 26(10), May 2016.
[546] Al-Naji, Ali, Perera, Asanka & Chahl, Javaan, "Remote Monitoring of Cardiorespiratory Signals from a Hovering Unmanned Aerial Vehicle", *BioMedical Engineering Online* 16(101), August 2017.
[547] Gevaert, Caoline, Sliuzas, Richar, Persello, Claudio & Vosselman, George, "Evaluating the Societal Impact of Using Drones to Support Urban Upgrading Projects", *International Journal of Geo-Information*, March 2018.
[548] Lidynia, Chantal, Philipsen, Ralf & Ziefle, Martina, "Droning on About Drones—Acceptance of and Perceived Barriers to Drones in Civil Usage Contexts", *Advances in Human Factors in Robots and Unmanned Systems*, 2017.
[549] Dempsey, 2015, op. cit.
[550] Hodgson, 2016, op. cit.
[551] Wilson, 2014, op. cit.
[552] Dempsey, 2015, op. cit.
[553] Hopkins, Anne, "The Ethical Debate on Drones", *Digital Commons*, 2017.
[554] Stansbury, Richard, Olds, Joshua & Coyle, Eric, "Ethical Concerns of Unmanned and Autonomous Systems in Engineering Programs", *121st ASEE Annual Conference & Exposition*, June 2014.
[555] Wilson, 2014, op. cit.
[556] Dempsey, 2015, op. cit.
[557] Al-Naji, et al., 2017, op. cit.

Two more topics that heavily play into responsibility and privacy is that of security and transparency.[558,559] Firstly, who is in control of the data or media collected by UAVs? Is anyone in possession of a drone entitled to the content it records? Are individuals allowed to publish this material, police allowed to make arrests based on it, or corporations able to use them to improve marketing? Further, when it comes to transparency, are individuals entitled to viewing data and footage generated by public-use drones like municipal drones or traffic-monitoring drones? If individuals are not allowed to even view this data, it would be hard to imagine being able to request deletion or 'opt-out' of UAV surveillance—especially if surrounding houses/properties do not. As such, issuance of data control, compliance, and security are also of high interest to this topic. Especially when it comes to warfare or defence applications such as UAV threat analysis, killings, and aerial support, maintaining strict security protocols and protections is paramount. Not does this help to ensure the correct individuals are being targeted, but also it aims to avoid friendly-fire or unauthorized use and access to military UAV capabilities that soldiers on the ground depend on for information and assistance.[560,561,562]

6.2.4. Self-driving vehicles

Self-Driving Vehicles (SDVs), more commonly known as "autonomous vehicles", raise many of the same ethical issues as UAVs do. The ethics of SDVs is a field that has attracted ample scholarly attention, with some preliminary discussions even dating back to 2010.[563] Standard issues in these discussions, even still today, are questions of security (e.g., how "hackable" vehicles are), responsibility (e.g., Does collision responsibility fall on the manufacturer, system programmer, or user?), and safety (e.g., Are SDVs *actually* safer than human drivers in all contexts?). Especially in safety discussions, one exceedingly popular topic is "ethical crashing".[564] Less commonly discussed (but still important) topics in SDVs are those of privacy, and data transparency and control.[565]

The most recent debates in SDV literature (no later than 2016) focus heavily on "ethical crashing". This is the term for the decision-making model SDVs ought to be programmed with for how to handle crashes when situations occur where a crash is unavoidable. Most discussions focus on the extreme cases—when cars need to choose between killing pedestrians, other drivers, or animals, and killing the car occupants themselves. These situations are frequently modelled on the trolley problem, and authors frequently use consequentialism, deontology or virtue ethics to assess the "right" course of action. As more researchers have approached ethical crashing by treating it as a trolley problem, many have found that the approach falls short and, at best, fails to provide guidance on most normal driving situations and contexts.[566] At worst, trolley models for SDVs end up causing authors to oversimplify

[558] Stansbury, 2014, op. cit.
[559] Gevaert, et. al., 2018, op. cit.
[560] Wilson, 2014, op. cit.
[561] Stansbury, 2014, op. cit.
[562] Manjikian, Mary, "A Typology of Arguments About Drone Ethics", *Strategic Studies Institute US Army War College*, October 2017.
[563] Steinfeld, Aaron, "Ethics and Policy Implications for Inclusive Intelligent Transportation Systems", *Carnegie Mellon Robotics Institute*, 2010.
[564] Thornton, Sarah, Pan Selina, Erlien, Stephen & Gerdes, Christian, "Incorporating Ethical Considerations Into Automated Vehicle Control", *IEEE Transactions on Intelligent Transportation Systems* 18(6), June 2017.
[565] Trimmer, Jelte, Pel, Bonno & Kool, Linda, et al., "5.3 Big Data", *Converging Roads: Linking Self-Driving Cars to Public Goals,* February 2015.
[566] Danks, David & London, Alex, "Regulating Autonomous Systems: Beyond Standards", *IEEE Intelligent Systems*, 2017.

the decisions being made by cars and mask other important issues by programming ethics based only on possible “worst case scenarios”. In their place, many of these authors are arguing for risk assessment models. How these are formulated differ slightly, but for this report, it is enough to point out that the trolley model is falling out of favour as an adequate way of programming vehicles, opening the door for further ethical debate on what to replace it with. Further, ethical discussions of the future will need to focus more on “mundane” cases of SDV decisions.[567,568,569,570]

Another unorthodox ethical consideration building from the revolt against trolley models is that of SDV customization. Some authors propose that individuals should be given the opportunity to choose their vehicle’s “moral programming” so that it is reflective of their own values. For example, the car’s programming could be set to allow individuals to choose self-preserving or self-sacrificing value profiles, rather than making a “one size fits all” vehicle that pulls more paternalistic in its decisions (i.e., a vehicle that makes the calls designers think are the “best” for all).[571,572,573,574] Aside from setting some unchangeable parameters (no vehicular homicide attempts or unnecessarily driving into oncoming traffic), a more customizable approach could help address two main issues facing adoption of SDVs: trust and value incongruity.

Trust in SDVs among users is achieved when they are confident that their car is making the “right” choice (in this case, “right” means the same choice that they would make if they were driving). A popular way to achieve this, as suggested by one author, is to make the decision-making processes of cars as transparent as possible, and utilizing democratic and participatory means to achieve consensus on a baseline for acceptable SDV decisions (i.e., self-sacrificing or self-preserving in nature).[575,576,577] Questions on *how* to achieve such a consensus, if trust is a desirable value to pursue in SDV use, and if such a broad ethical framework should be applied to SDVs are all ethical queries of import here. Why trust is so difficult to achieve and such a big barrier to SDV adoption is that of value incongruity. In short, value incongruity can occur when users want different decision-making models for themselves than for other drivers. For example, a user may want self-preserving values to be applied to their own vehicle, but think vehicles of other drivers should utilize utilitarian values,[578] or different cultures desire cars with different values[579], or some individuals refusing to use SDVs but want everyone else to adopt

[567] Etzioni, Amitai & Etzioni, Oren, “Incorporating Ethics into Artificial Intelligence”, *The Journal of Ethics* 21(4), December 2017.
[568] Goodall, Noah, “From Trolleys to Risk: Models for Ethical Autonomous Driving”, *American Journal of Public Healthy*, April 2017.
[569] Himmelreich, Johannes, “The Everyday Ethical Challenges of Self-Driving Cars”, *The Conversation*, March 2018.
[570] Shariff, Azim & Rahwan, Iyad, “Psychological Roadblocks to the Adoption of Self-Driving Vehicles”, *Nature: Human Behaviour*, September 2017.
[571] Danks, et. al., 2017, op. cit.
[572] Applin, Sally, “Autonomous Vehicles Ethics, Stock or Custom?”, *IEEE Consumer Electronics* 6(3), June 2017.
[573] Keeling, Geoff, “Legal Necessity, Pareto Efficiency & Justified Killing in Autonomous Vehicle Collisions”, *Ethical Theory and Moral Practice* 21(2), April 2018.
[574] Wamsley, Laurel, “Should Self-Driving Cars Have Ethics?”, *Technology*, October 2018.
[575] Shariff, et al., 2017, op. cit.
[576] Applin, 2018, op. cit.
[577] Lin, Patrick, “Why Ethics Matters for Autonomous Cars”, *Autonomous Driving: Technical, Legal and Social Aspects*, 2016.
[578] Shariff, et al., 2017, op. cit.
[579] Maxmen, Amy, “Self-Driving Car Dilemmas Reveal that Moral Choices are Not Universal”, *Nature*, October 2018.

them.[580] These issues add fuel to the moral fire of whether SDVs should be universally programmed or not.

Further ethical concerns relate to the potential of stereotyping or biases existing in the data used for decision making in SDVs,[581] distributive justice,[582] and determining the safest, least-disruptive approach to phase-in SDVs on a large scale.[583] Finally, it is worth noting that some of the ethical issues researchers have been focusing on may be irrelevant given substantial changes to current infrastructure, public policy, traffic management practices alongside (or before) the introduction of SDVs. Sky bridges for pedestrians, underground car-only roads, speed limits may make a substantial number of the ethics issues discussed in the literature moot.[584]

6.2.5. Telerobotic systems

Telerobotics, i.e., semi-autonomous robots operated from a distance, are used in many different fields, especially in the healthcare and military sectors.[585] As they have a human agent operating them either through wireless or wired communication, they do not raise the same ethical issues that the autonomous robots discussed in this report.[586] They also have existed for longer than autonomous robots. Nonetheless, research in this area continues to develop and new ethical issues are emerging. Ethical issues raised depend to a large extent on the field of application in which they are used.

A key ethical issue that telerobotics raises and that has impacts on the different fields of application where such technologies are used relates to the distance in human relationships that this technology generates. For instance, In the healthcare sector telerobotics systems make it possible for a healthcare practitioner to provide care at a distance from the patient. This distance in and by itself creates particular issues. While it can be advantageous, e.g., reducing need to commute, lesser costs, faster access, it might threaten the relationship of face-to-face and the experience of touch-based care that is a core dimension of healthcare, and the trust that is essential to this practice.[587] Furthermore, it might lead to a care that is less personalised, which is also a central element of healthcare.[588]

This distancing is also evident in the military sector and raises some ethical concerns. In particular, one of the most morally concerning aspects that touches this sector relates to the ability to activate weapons at a distance. In the case of the use of military drones, Asaro talks about "bureaucratised killing" and shows that this technology "presents far more potential targets and shapes the interpretations and determinations of targets in unpredictable ways."[589] Military personnel might also

[580] Sparrow, Robert & Howard, Mark, "When Human Beings are Like Drunk Robots: Driverless Vehicles, Ethics, and the Future of Transport", *Transportation Research*, May 2017.

[581] Lin, 2016, op. cit.

[582] Goodall, 2017, op. cit.

[583] Sparrow, 2017, op. cit.

[584] Himmelreich, 2018, op. cit.

[585] Avgousti, Sotiri, et al., "Medical telerobotic systems: current status and future trends", *Biomedical Engineering OnLine*, Vol. 15, No. 96, 2016. Sullins, John P., "Introduction: Open Questions in Robotics", *Philosophy & Technology*, Vol. 24, 2011, pp. 233-238.

[586] Sullins, John P., "Introduction: Open Questions in Robotics", *Philosophy & Technology*, Vol. 24, 2011, pp. 233-238.

[587] Worms, Frédéric, "The Two Concepts of Care. Life, Medicine, and Moral Relations." *Esprit*, No. 1, Jan 2006, pp. 141-156.

[588] Ibidi; Avgousti, et al., op. cit., 2016, p. 34.

[589] Asaro, Peter. W., "The labor of surveillance and bureaucratized killing: new subjectivities of military drone operators", *Social semiotics*, Vol. *23*, No. 2, 2013, p. 220.

be psychologically affected by remote killing, i.e., away from the battlefield.[590] It depersonalises such actions and might have adverse effects on whether people feel responsible and/or accountable for their actions. As such, telerobotics change the conduct of warfare in morally concerning ways.

Another set of ethical issues that telerobotics raises is the increased technologisation it enables. This affects particularly the healthcare sector in which it accompanies the development of a particular type of medicine, i.e., one that is highly technological and that tends to treat the body as a machine made of different parts that can be treated separately.[591] This is done at the expense of other forms of medicines that might be more holistic, singularised, and less technological.

Another ethical issue related to the type of medicine that is promoted by the use of telerobotics in the healthcare sector relates to the high costs to develop, acquire, implement, and maintain the technology.[592] In turn, the high cost of the technology might further increase inequality in healthcare, between those who can afford it and those who cannot. This might further entrench healthcare inequalities within Europe as well as between the Global North and the Global South. It is important to note that promoting research in such a highly technological form of medicine is done at the expense of other forms of medicine that could benefit many more people and to which many more people might have easier access.

Another set of ethical issues of telerobotics relates to "wireless networks security vulnerability".[593] Here as well, the healthcare sector clearly demonstrates the risks at stake. In telemedicine, i.e., medicine conducted at a distance, this vulnerability raises "major concern for the exploitation of (long-distance) telerobotics".[594] Evans *et al*. point to this risk and the way it is significantly impeding implementation of telerobotics in this area. They explain: "Long-distance telerobotics require fast and reliable data connections capable of transmitting a large quantity of data. […] Stable networks are often not available in remote geographical locations, the same areas that could benefit most telerobotics."[595] This point questions the argument that is often made to support the development of healthcare telerobotics and telemedicine, i.e., that they will make it possible to bring expert care to areas where such expertise is lacking.[596] On the contrary, wireless networks vulnerability may lead to further increasing already existing inequalities in healthcare. Cybersecurity vulnerability also creates privacy and confidentiality issues. If networks are not sufficiently secured, there is the risk of leakage of highly sensitive private information, which is what healthcare data is.[597] The hacking of such systems or its malicious use in any other manner may have potentially deeply harmful or fatal consequences for patients.[598]

Liability and responsibility issues are also created by the use of telerobotics.[599] In case of a complication following a surgery for instance, who is to be held responsible? The telesurgeon, the designer of the telerobot, the provider/supplier, or the entity that certified/approved it?

[590] Asaro, op. cit., 2013.
[591] Worms, op. cit., 2006.
[592] Avgousti, et al., op. cit., 2016, p. 34; Evans, Chadrick R., et al., "Telemedicine and telerobotics: from science fiction to reality", *Updates in Surgery*, Vol. 70, 2018, p. 361.
[593] Avgousti, et al., op. cit., 2016, p. 34.
[594] Avgousti, et al., op. cit., 2016, p. 34.
[595] Evans, et al., op. cit., 2018, p. 361.
[596] Gallego, Jelor, "The Microbots Will Treat Diseases From Inside Your Body", 9 Oct 2016. https://futurism.com/meet-the-microbots-that-will-treat-diseases-from-inside-your-body
[597] Evans, et al., op. cit., 2018, p. 361.
[598] Ibid.
[599] Avgousti, Sotiri, et al., op. cit., 2016, p. 34.

6.2.6. Robotic exoskeletons

Robotic exoskeletons refer to a class of mechanized, wearable robotics that enhance the physical capabilities of the wearer. Robotic exoskeletons are also, in some discussions, considered to be a part of the "collaborative robots" domain, but will be treated here separately as they are atypical representations of co-bots. Unfortunately, the ethical debate on robotic exoskeletons is sparse and underdeveloped—issues were first noted in 2014, but have since then garnered relatively little attention.[600,601]

One of the topics that generates the most concern in robotic exoskeletons is that of accessibility. Researchers maintain concerns about wealth distribution, especially when discussing contexts of healthcare or consumer market use. Only allowing individuals who can afford this technology to utilize it may cause significant socio-economic consequences in the form of stereotyping and discrimination (more will be said about this further on).[602] Further issues in accessibility may be seen in that of maintenance and repairs, with individuals living in more rural or less developed areas unable to benefit from robotic exoskeletons due to the lack of facilities to keep them properly functioning and performing repairs. Theses current factors have robotic exoskeletons appearing to be a technology of privilege, rather than one of enabling social equality.[603,604]

Further concerns develop in the realm of addiction— if robotic exoskeletons enable those with physical disabilities to live the lives they want to better and with less pain, then it may be the case that these users become heavily dependent on this technology. They would view themselves and their personal situation as being "worse off" without a robotic exoskeleton than with it. Not only this, but users may potentially forget or lose their ability to function independently of the exoskeleton—the de-learning of fine motor skills and the erosion of muscle tissue being two such examples. Thus, there may be the potential for withdrawal if the exoskeleton breaks down or malfunctions or the user, for whatever reason, is unable to continue using an exoskeleton. As such, it would be necessary that more research be done to this end before heavy usage of robotic exoskeletons in every-day use contexts occurs to avoid negative psychological and physiological impacts.[605,606] Bridging on to this problem is the dilemma of ableism—where society may grow to see individuals who do not use robotic exoskeletons more negatively disabled or helpless than those who do. Especially problematic if users are unable to acquire exoskeletons due to cost or location and for individuals with disabilities that are not a good fit for exoskeleton use. This problem could be perpetuated if disabled persons are *expected* to use exoskeletons, so that the design of infrastructure no longer caters to individuals in wheelchairs, crutches, or scooters.[607,608,609]

[600] Sadowski, Jathan, "Exoskeletons in a Disabilities Context: The Need for Social and Ethical Research", *Journal of Responsible Innovation*, May 2014.
[601] Greenbaum, Dov, "Ethical, Legal and Social Concerns Relating to Exoskeletons", *Computers and Society* 45(3), September 2015.
[602] Bissolotti, Luciano, Nicoli, Federico & Picozzi, Mario, "Domestic Use of the Exoskeleton for Gait Training in Patients with Spinal Cord Injuries: Ethical Dilemmas in Clinical Practice", *Frontiers in Neuroscience*, February 2018.
[603] Sadowski, 2014, op. cit.
[604] Solinsky, Ryan & Specker Sullivan, Laura, "Ethical Issues Surrounding a New Generation of Neuroprostheses for Patients with Spinal Cord Injuries", *PM&R* 10(9), September 2018.
[605] Sadowski, 2014, op. cit.
[606] Greenbaum, 2015, op. cit.
[607] Sadowski, 2014, op. cit.
[608] Greenbaum, 2015, op. cit.
[609] Solinsky, and Specker, 2018, op. cit.

Outside of healthcare and commercial applications are concerns surrounding manufacturing and industrial use contexts. Some researchers express concerns that the widespread use robotic exoskeletons in these fields may lead to the dehumanization or overworking of industrial labourers—as the expectations of labour rise with the capabilities, but the compensation and working hours do not change.[610] Further concerns in this area surround use requirements: will exoskeletons be mandatory for certain places of employment? Will their use be optional? Or will they be required for some jobs like certain gloves or clothing materials?[611]

A more general issue that overlaps all use contexts of robotic exoskeletons is that of privacy. It is uncertain what types of data will be collected from exoskeleton users, but it is important that the data collection be transparent, able to be opted-in to, and users have control over its storage, deletion, and use.[612] On the design of these products, there is limited commentary, but the article that discusses the design of robotic exoskeletons asserts the need for an ethical design framework and suggests the implementation of one that is proactive, value-sensitive, and highly participatory in ensuring the best chances at ethical adoption and user outcomes.[613] Alongside this, safety and security measures involving the use and implementation of exoskeletons are also in critical need for development to ensure such decisions are not made, and potentially abused, on a corporate level.

6.2.7. Biohybrid robots

Also known as biomimicry systems, biohybrid systems, and bio-bots, biohybrid robots can refer both to organically grown and mechanically constructed components or to robots that imitate key features of organic entities (mobility, function, etc.). Dominantly biohybrid robots refer to robots created from human or animal cells, but the less-popular, but steadily growing subcategory *flora robotica* utilizes plant cells for construction instead (*flora robotica* will be addressed near the end of this section).[614] Given the novelty of this field, it is difficult to assess where ethical areas of import will be. Webster-Wood, one of the leading scientists in biohybrid robotics recommends using existing policies and ethical guidelines in synthetic biology as a starting point until biohybrid robotics develops further.[615] Bearing this guidance in mind, there are a few potential problems to keep in mind as the field of biohybrid robotics develops, one of these precautionary concerns is that of the 'emergent behaviour' problem.[616] As of yet, researchers are still struggling to gain control of these, even small-scale, biohybrid robots.[617,618] Thus, while one part of the biohybrid model may operate as intended, it may change entirely when combined with other parts of a biohybrid system— resulting in steep consequences, especially if researchers are unable to regain control of the robot due to size,

[610] Greenbaum, 2015, op. cit.

[611] Maurice, Pauline, Allienne, Ludivine & Malaise, Adrien et al, "Ethical and Social Considerations for the Introduction of Human-Centered Technologies at Work", *IEEE Workshop on Advanced Robotics and its Social Impacts*, June 2018.

[612] Greenbaum, 2015, op. cit.

[613] Lenca, Marcello, Kressig, Reto & Jotterand, Fabrice et al., "Proactive Ethical Design for Neuroengineering, Assistive and Rehabilitation Technologies: The Cybathlon Lesson", Journal of Neuroengineering *Rehabilitation* 14, November 2017.

[614] Webster-Wood, Victoria, Akkus, Ozan & Gurkan, Umut et al. "Organismal Engineering: Toward a Robotic Taxonomic Key for Devices Using Organic Materials", *Science Robotics Review*, November 2017.

[615] Webster-Wood, et al., 2017, op. cit.

[616] Ibid.

[617] Webster-Wood, Victoria, "Biohybrid Robots Built from Living Tissue Start to Take Shape", *The Conversation*, August 2016.

[618] Zhang, Chuang, Wang, Wenxue & Xi, Ning, et al., "Development and Future Challenges of Bio-Syncretic Robots", *Research Robotics Review*, August 2018.

complexity, or other variables. It is important that biohybrid research works heavily with ethicists and policy makers to ensure success and avoidance of cross boundaries that may be difficult to ethically uncross once decided upon.[619,620]

Aside from these ethical issues researchers have self-identified, a few other issues may be worth investigating, since many of these approaches are being used for biomimicry of animals,[621,622] one potential concern could be that of organic-synthetic organism relations— making sure animals are not overly stressed by the machines and questions of reproduction and relation would also be of concern—if an animal that mates for life, like a sea horse, beaver, or gibbon, decides upon a synthetic mate, is it ethical to allow this to happen if the synthetic creature is unable to reproduce or is only there for a short time as an experiment? The emotional distress of such situations may be unethical to infringe upon animals.

Further, if biohybrid robots are given bodies that have enhanced sensory, response, and mobility capabilities, questions of sentience may come to the fourfold of robotics discussions unlike ever before—as robots could very well experience pain in biohybrid systems.[623,624] Fuelling the bigger fire on robot rights and acceptable treatment of robots very quickly. While these researchers may be excited that biohybrid robots will be more sensible, tactile interactors than what their current metallic materials allow, but this change in bodily experience may carry with it more moral problems than policy currently accounts. It may be hard to speculate on this topic, but it seems fair to assess that the design of living, feeling creatures, of biohybrid construction or not, may change ethical guidelines for treatment. Additional concerns may surface in the biomedical sphere if robots are being used to grow human tissue for harvesting later.[625] While it may be something of not immediate concern at the current stage of research, it becomes concerning to think humans may be growing new creatures purely for the purpose of using them as a means to heal ourselves. Even if it is only morally suspect and not of any direct wrongdoing, it certainly seems like a peculiar precedent to set.

Turning towards *flora robotica,* this field will likely generate more commentary as it becomes more popular, but for now and the near future, there seem to be few major ethical concerns. If one were to speculate on topics of ethical intrigue, many of these could heavily depend on one's attitude towards the moral status of plants or bio-fabrication in general; some arguments for plant welfare and wellbeing, creating feeling experiments, and the "rightness" of human-controlled bio-structural modifications may be able to be made. However, it is far too early to see which of these arguments

[619] Webster-Wood, et al., 2017, op. cit.

[620] Raman, Ritu & Bashir, Rashid, "Biomimicry, Biofabrication, and Biohybrid Systems: The Emergence and Evolution of Biological Design", *Advanced Healthcare Materials*, 2017.

[621] Romano, Donato, Donati, Elisa, Benelli, Giovanni & Stefanini, Cesare, "A Review on Animal-Robot Interaction: From Bio-Hybrid Organisms to Mixed Societies", *Biological Cybernetics*, October 2017.

[622] Mariano, Pedro, Salem, Ziad & Mills, Rob et al., "Design Choice for Adapting Bio-Hybrid Systems with Evolutionary Computation", *GECCO 2017 Companion*, July 2017.

[623] Coyle, Stephen, Majidi, Carmel, LeDuc, Philip & Hsia, Jimmy, "Bio-Inspired Soft Robotics: Material Selection, Actuation, and Design", *Extreme Mechanics* 22, July 2018.

[624] UCLA Smueli Newsroom*,* "UCLA Bioengineering Leads Development of Stingray-Inspired Soft Biobot", January 2018.

[625] Mouthuy, Pierre-Alexis & Carr, Andrew, "Growing Tissue Grafts on Humanoid Robots: A Future Strategy in Regenerative Medicine?" *Science Robotics* 2(4)*,* March 2017.

will hold to be worthwhile exploring—there is much research to be done. Please refer to the footnotes for some ongoing research on *flora robotica* for understanding the ongoing progress in the field.[626,627]

Finally, a reoccurring problem that biohybrid robotics has not avoided, despite being unmentioned, is that of waste management practices—bio-waste and e-waste will both see increases if these types of robotics grow to be commonplace, and it would be important to have a recycling or disposal system ready.

6.2.8. Swarm robots

Swarm robots, also called "collective robots" or "distributed collaborative systems" are systems that "demonstrate collective decision-making without human help".[628] They are one of the key emerging fields of robotics research today and are attracting much attention, especially in the military sector, disaster response, and space exploration. Instead of human beings, they can enter dangerous areas (whether in wars or disaster settings for instance) and avoid loss of life and expensive equipment (as individual robots of a swarm are generally simple and inexpensive).[629] However, they also raise a number of ethical issues that this section identifies. This section begins by highlighting a set of issues that arise with such robots, i.e., privacy and surveillance, risk of hacking, and environmental costs. It also points to the ethical risks created by the use of this technology in the military sector. It concludes with more fundamental conceptual, ontological, and ethical considerations that swarm robots raise.

One of the strengths of swarm robots consists in their highly adaptive nature: they can adapt to any environment, especially changing ones. However, this makes them also particularly unpredictable and therefore leads to questions of responsibility and accountability. As Singer puts it, "[s]warms may not be predictable to the enemy, but neither are they exactly controllable or predictable for the side using them, which can lead to unexpected results: [...] a swarm takes action on its own."[630] This technology has great surveillance power and this raises deep privacy issues. This risk is further exacerbated when swarm robots are designed to be small or invisible or in a way that enables them to covertly penetrate any area.[631] Furthermore, the decentralised nature of the technology makes it particularly resilient as the destruction of one component does not mean the destruction of the whole system. This makes this technology even more robustly intrusive, and therefore, a potential threat to privacy.[632] An

[626] Hamann, Heiko, Divband Soorati, Mohammad & Heinrich, Mary Katherine et al., "Flora Robotica— An Architectural System Combining Living Natural Plants and Distributed Robots", *Cornell University Computer Science & Emerging Technologies*, September 2017.

[627] Skrzypczak, Tomasz, Krela, Rafal & Kwiatkowski, Wojciech et al, "Plant Science View on Biohybrid Development", *Frontiers in Bioengineering and Biotechnology*, 2017.

[628] Lachow, Irving, "The Upside and Downside of Swarming Drones," *Bulletin of the Atomic Scientists*, Vol. 73, No. 2, 2017, p. 96; Bredeche, Nicolas, Haasdijk, Evert, and Prieto, Abraham, "Embodied Evolution in Collective Robotics: A Review," *Frontiers in Robotics and AI*, Vol. 5, No. 12, 2018; Magnuson, Stew, "Military Beefs Up Research Into Swarming Drones," *National Defense Magazine*, March 1, 2016. https://www.nationaldefensemagazine.org/articles/2016/2/29/2016march-military-beefs-up-research-into-swarming-drones

[629] Lachow, op. cit., 2017, p. 96; David Grémillet et al., "Robots in Ecology: Welcome to the Machine," *Open Journal of Ecology*, Vol. 2, No. 2, 2012, p. 54; Scharre, Paul, "Robotics on the Battlefield Part II. The Coming Swarm", Center for a New American Security, October 2014, p. 5–6; Magnuson, op. cit, 2016.

[630] Singer quoted in Mark Coeckelbergh, "From Killer Machines to Doctrines and Swarms, or Why Ethics of Military Robotics Is Not (Necessarily) About Robots," *Philosophy & Technology*, Vol. 24, 2011, p. 273.

[631] See for instance the swarm robots developed by engineers at the University of Harvard. Programmable Robot Swarms, Wyss Institute, University of Harvard. https://wyss.harvard.edu/technology/programmable-robot-swarms/

[632] Roff quoted in Lachow, op. cit., 2017, p. 96.

additional ethical issue that arises with this technology relates to the risk of hacking and its high dual use potential that could have significant impacts on human life and society.[633] Another ethical issue relates to their environmental cost, especially "the end of that product lifecycle".[634] As Lin observes, "[t]hey may contain hazardous materials, like mercury or other chemicals in their battery, that can leak into the environment. Not just on land, but we also need to think about underwater and even space environments, at least with respect to space litter."[635] As this technology gets wider use, such effects would increase. The use of swarm robots in the military sector also raises ethical concerns.[636] In particular, the faster reactions rendered possible by this technology might lead to an increased risk of quick escalation in military conflict and, eventually, "make it easier to start a war."[637]

Beyond the practical and concrete ethical issues that swarm robots raise, it is essential to point to more fundamental ethical issues that they have the potential to create due to the high degree of autonomy, adaptability, and resilience that they exhibit. As Bredeche, Haasdijk, and Prieto note, swarm robots are characterised by an "autonomy that occurs at two levels: not only the robots perform their tasks without external control but also they assess and adapt—through evolution—their behaviour without referral to external oversight and so learn autonomously. This adaptive capability allows robots to be deployed in situations that cannot be accurately modelled *a priori*."[638] As such, these robots push one step further the emancipation of the technology from the human creator. In turn, this raises ethical tensions that are, for the moment, insolvable. This tension is exemplified by the position of Bredeche Haasdijk, and Prieto on this technology. While on the one hand they claim that we should keep a human in the loop when it comes to swarm robots, on the other hand, they want to design these robots to be the most autonomous possible and, hence defer responsibility to the machine itself.[639] These are two contradictory positions that policy-makers, regulators and society will eventually have to decide upon. Coeckelbergh identifies such systems as "cloudy and unpredictable systems, which rely on decentralized control and buzz across many spheres of human activity."[640] As he demonstrates, swarm robots question classical ethical frameworks founded on an ontology of technology as tools created by humans.[641] In turn, this challenges "the assumptions of our traditional theories or responsibility."[642] Eventually, swarm robots bring us one step closer to the classic science-fi scenario of machines emancipated from their human creator and the danger that robots take control over humanity. This is even more worrying as this technology is developed in the military sector and could, in the future, be further equipped with weapons. Furthermore, the possible prospect of swarm robots reproducing themselves autonomously through 3D printing makes this concern even starker.[643]

[633] Lachow, op. cit., 2017, p. 98.

[634] Lin, Patrick, "Drone-Ethics Briefings: What a Leading Robot Expert Told the CIA," *The Atlantic*, December 21, 2011.

[635] Ibid.

[636] John Arquilla and David Ronfeldt, "Swarming & The Future of Conflict", RAND, National Defense Research Institute, 2000. See also section on use of AI and robotics in the defence sector in the present report.

[637] See Asaro's work quoted in Coeckelbergh, op. cit., 2011, p. 271; Scharre, October 2014, p. 7.

[638] Bredeche, Haasdijk, and Prieto, op. cit. 2018, p. 1.

[639] Ibid, p. 12.

[640] Coeckelbergh, op. cit., 2011, p. 269.

[641] Ibid, p. 274.

[642] Ibid.

[643] Bredeche, Haasdijk, and Prieto, op. cit. 2018, p. 12.

6.2.9. Microrobots

Microbots, or micro-robots, are used to access hard-to-reach areas, such as environments that are too dangerous or too small for humans or other bigger robots. There is particular interest for them in the medical[644] and the military sectors.[645] Their main added value consists in their being small and cheap. It is mainly when they are associated to other microbots that they gain particular power. Ethicists have primarily thus far focused their attention on swarm robots. Nonetheless, even independently from collective behaviour, microbots raise ethical issues that this section seeks to identify.

To begin with, their small size makes them potentially highly intrusive, whether in the human body or in communities. For instance, they may be inserted in the body of a patient and controlled remotely.[646] This raises privacy issues.[647] If microbots are equipped with surveillance technology, this might also be ethically problematic as gives them the capacity to carry out covert surveillance, hence further expanding surveillance over individuals and society. Furthermore, they raise issues of control and ownership, especially when inserted in the human body.

In addition, considering that they are inexpensive, losing them is not considered to be an issue. Users might therefore be even more inclined to send them to hard-to-reach areas as the cost of losing them is relatively low. However, this might have environmental costs (e.g., if non-biodegradable) and create potential hazards (environmental waste, for example). For instance, Lin notes: "How do we clean up after them? If we don't, and they're tiny--for instance, nanosensors--then they could then be ingested or inhaled by animals or people. […] They may contain hazardous materials, like mercury or other chemicals in their battery, that can leak into the environment. Not just on land, but we also need to think about underwater and even space environments".[648] Furthermore, their insertion in the human body can lead to dramatic outcomes should human control over them be lost or if they are hacked.

Finally, when used in the military sector, Kladitis notes that microbots or nanobots could be perceived as chemical or biological weapons.[649] This raises regulatory issues that will need to be addressed.

6.2.10. Collaborative robots

Collaborative robots, frequently shortened to cobots, refer to a class of robots whose primary focus is to perform tasks in tandem with human labourers. Some examples of this can be robotic arms holding skin and handing the surgeon tools during surgery, a co-bot that welds metal pieces together as a human places the pieces, or a robot that lifts and moves heavy objects for its human co-workers. No matter the context or application, one of the critical concerns surrounding the use of cobots is that of trust and psychological harm for human co-workers. When workplaces transition to these "Industry

[644] Freeman, Tami, "Magnetic microbots line up for stem cell therapy", *Physics World*, 30 May 2019. https://physicsworld.com/a/magnetic-microrobots-line-up-for-stem-cell-therapy/; Gorey, Colm, "Tiny robots in our blood could soon be used to sniff out and treat cancer", *Silicon Republic*, 23 Nov 2017. https://www.siliconrepublic.com/machines/blood-cell-sized-robots-cancer; Capgemini, "Microbots: Innovation in Healthcare", 2 Dec 2014. https://www.capgemini.com/2014/12/microbots-innovation-in-healthcare-0/.

[645] Kladitis, Paul E., "How small is too small? True microrobots and nanorobots for military applications in 2035", Research Report, Maxwell Air Force Base, Alabama, April 2010 and see also references in the section on Swarm robots.

[646] Gallego, Jelor, "The Microbots Will Treat Diseases From Inside Your Body", 9 Oct 2016. https://futurism.com/meet-the-microbots-that-will-treat-diseases-from-inside-your-body

[647] Roff quoted in Lachow, "The Upside and Downside of Swarming Drones," *Bulletin of the Atomic Scientists*, Vol. 73, No. 2, 2017, p. 96

[648] Ibid.

[649] Kladitis, Paul E., op. cit., April 2010, pp. 44-45.

4.0" spaces where robots and humans are required to work collaboratively, the transition is rarely gradual. This leads to a situation of forced adaptation for human labourers if they wish to keep their job: they must part with the traditional concept of a workplace and learn how to communicate and operate within a collaborative space.[650] This forced adaptation also means labour upskilling, which may inadvertently force older labourers out of the workplace, as they are unable to keep up with the necessary work and risk posing safety risks to themselves and other co-workers. In part, upskilling is technical, but it is also psychological as the need to accept and trust workplace cobots is paramount in these spaces.[651] Cobot workers need to learn how to effectively communicate with cobots and be taught the limitations and functions of the cobot so that labourers are not deceived into thinking the cobot can take instructions and adapt as a human co-worker can.[652]

Undoubtedly, increased human-robot interactions demand increasing dependency upon the robot to function correctly. Thus, concerns are not only related to a greater need for training and implementation management solutions, but also to different regulations and oversight on maintenance and repairs. For situations in which cobots do malfunction or miscommunication occurs, there is a greater need for modified liability and responsibility regulations so that legal action may be a form of recourse for injured workers. As it stands now, it is difficult to tell if problems occur due to a technical flaw, lack of training, miscommunication, or lack of maintenance and who is liable: the worker, the robot, the company who owns it, the robots' creators?[653]

Finally, even cobots cannot escape the problems of privacy and security. As more sensors and data will need to be collected from its surroundings and human co-workers in order to have the necessary flexibility and responsiveness to be safe, more questions about user privacy and data retention arise. Further, security measures need to be implemented to avoid both in-house problems and external interference. Two-step verification for tasks and authentication for commands may be a necessity in avoiding honest mistakes and misunderstandings in giving cobots commands and allowing manipulation or theft of worker data. While some control and flexibility of the cobot's behaviour is necessary for adoption in a wide variety of workplaces, some capabilities may be better left black-boxed (gesture recognition changes, critical operative functions, speed) for general workers and only able to be manipulated by designated personnel.[654]

[650] Bendel, Oliver, "Co-Robots from an Ethical Perspective", *Business Information Systems and Technology 4.0*, March 2018.

[651] Buoncompagni, Luca, Capitanelli, Alessio, & Carfi, Alessandro et al., "From Collaborative Robots to Work Mates: A New Perspective on Human-Robot Cooperation", *ERCIM News*, July 2018.

[652] Salem, Maha & Dautenhahn, Kerstin, "Evaluating Trust and Safety in HRI: Practical Issues and Ethical Challenges", *The Emerging Policy of Ethics of Human Robot Interaction,* March 2015.

[653] Maurice, Pauline, Allienne, Ludivine, & Malaise, Adrien et al., "Ethical and Social Considerations for the Introduction of Human-Centered Technologies at Work", *IEEE Workshop on Advanced Robotics and its Social Impacts,* 2018.

[654] Fletcher, Sarah R, and Phil Webb, "Industrial Robot Ethics: Facing the Challenges of Human-Robot Collaboration in Future Manufacturing Systems", *A World with Robots: International Conference on Robot Ethics: ICRE 2015*, 2017.

7. Ethical analysis: Ethical issues in different AI and robotics application domains

In this section, we identify and describe the main ethical issues with regard to artificial intelligence and robotics technology *applications*. As stated in the methods section, in this ethical analysis, we follow the *Anticipatory Technology Ethics* approach developed by Brey (2012).[655] Having focused on the technology level and the artefact level in section 5 and section 6, we now turn our attention to the *application level* of the approach's three-level system of ethical analysis.

Our objects of analysis at this level consist of uses of the (previously identified) technological artefacts (or products) and procedures in particular domains or contexts, for particular purposes, and by particular user groups. Thus, in this section, we discuss the ethical issues that may occur in certain application domains of AI and robotics technology, such as *transportation*, *defence*, *healthcare*, and *finance and insurance*. In addition, we discuss issues with regard to products that are specific to particular application domains of AI and robotics. Furthermore, we detail specific ethical issues for different types of users of AI products and robotics products.

In this section, we again focus on both present issues and issues that may occur between now and 20 years into the future. Most of our analysis in this section is based on an extensive analysis of the academic and popular literature on ethical issues in AI and robotics applications. In addition, we have made use of the results of our SIENNA expert workshops and expert interviews, and we have on occasion used ethical checklists to conduct our own analysis in areas where the literature was sparse.

This section is structured as follows. Subsections 7.1 and 7.2 offer brief descriptions of the most important ethical issues that may present themselves in the main AI application domains and robotics application domains, respectively. In our descriptions, we put special emphasis on interesting and/or unique ethical issues. Then, subsections 7.3 details specific ethical issues for different types of users and stakeholders of AI and robotics technologies.

7.1. Ethical issues with AI applications

This subsection identifies and describes the ethical issues that may occur in various important application domains of AI technology. It discusses, in turn, the issues in *infrastructure and cities* (subsection 7.1.1), *healthcare* (subsection 7.1.2), *finance and insurance* (subsection 7.1.3), *defence* (subsection 7.1.4), *law enforcement* (subsection 7.1.5), the *legal sector* (subsection 7.1.6), *public services and governance* (subsection 7.1.7), *retail and marketing* (subsection 7.1.8), *media and entertainment* (subsection 7.1.9), *smart home* (subsection 7.1.10), *education and science* (subsection 7.1.11), *manufacturing* (subsection 7.1.12), and *agriculture* (subsection 7.1.13). Table 10 below lists the most important ethical issues that have been identified for each of these application domains.

Domain	**Ethical issues**	
Infrastructure & cities	**- Safety** **- Security** **- Privacy**	**- Freedom and autonomy** **- Trust**

[655] Brey, P.A.E., 2012, op cit.

Healthcare	- Privacy - Informed consent	- Discrimination & health inequality - Trust
Finance & insurance	- Safety - Security	- Bias and discrimination - Responsibility and accountability
Defence	- Just war compliance - Threat of uncontrolled escalation	- Responsibility and accountability
Law enforcement	- Bias and discrimination - Privacy	- Transparency and accountability
The legal sector	- Bias and discrimination - Responsibility and accountability	- Autonomy and freedom
Public services & governance	- Gov. distancing from citizens - Depersonalised services - Exacerbation of social inequality - Transparency and accountability	- Freedom - Privacy - Security - Politics and democracy
Retail & marketing	- Privacy - Autonomy - Bias and discrimination	- Community and wellbeing - Harms from inaccurate inferences
Media & entertainment	- Impoverished journalism - Diminished human creativity - Autonomy	- Privacy - Freedom (of speech) - Democracy
Smart home	- Privacy - Autonomy	- Bias - Exacerbation of social inequality
Education & science	- Quality of education - Bias - Transparency - Privacy	- Informed consent - Research integrity - Social responsibility
Manufacturing	- Job losses - Social inequality - Privacy - Autonomy	- Responsibility and accountability - Loss of diversity - De-skilling
Agriculture	- Power asymmetries - Industrial monocultures	

Table 10: Overview of ethical issues in major AI application domains.

7.1.1. Infrastructure & cities

AI technology may be used for infrastructure and cities to create a so-called "smart city". IBM describes a smart city as a city where its components and its citizens are instrumented, interconnected, and intelligent.[656] Instrumented implies that a city and its citizens are provided with infrastructures and devices that respond to a network. The information they provide to the network is then available to the other devices, making them interconnected. Analysing and using this information makes a city

[656] Elmaghraby, Adel S., and Michael M. Losavio, "Cyber security challenges in Smart Cities: Safety, security and privacy," *Journal of advanced research,* Vol. 5, No. 4, 2014, pp. 491-497., p. 492

intelligent. A smart city is a combination of different aspects, such as smart economy, smart mobility, smart environment, smart people, smart living and smart governance,[657] and aims to improve efficiency, safety, and convenience of its citizens by using smart technologies.[658] These technologies are applied to buildings and transportation systems, for example.

Currently, one of the main ethical discussions surrounding smart cities concerns cyber security and privacy. These concepts focus on "(1) [t]he 'privacy' and confidentiality of the information, (2) [t]he integrity and authenticity of the information, and (3) [t]he availability of the information for its uses and services."[659] Due to the interconnectedness of a smart city's constituent elements, separate data sets may be combined, thereby revealing sensitive data about citizens. This may lead to easier identification of individuals.[660] Furthermore, since the integrated data is likely to be stored in a cloud storage system, this raises question about who has access to the data, who is responsible for them, whether individuals can easily request removal of their personal data, and whether the data may be vulnerable to hacking and other malicious attacks. A major cyber security issue nowadays is Denial-of-Service (DoS) attacks. Such attacks block a connection on which services and devices are reliant and can cause physical harm. DoS directly affects physical safety when such an attack blocks (for example) hospital services, thereby putting patients' lives in danger.[661] Such threats also make DoS particularly suitable to be used for blackmail.

Secondly, as more and more components of a city are interconnected, this may increase potential harm to individual privacy and safety. For example, if a hacker can see whether someone's car is not parked near the house but instead driving around town, this makes his house more prone to burglary. Worse still, if such a GPS system were to be hacked, it could severely endanger individuals that are being stalked or are escaping (domestic) violence, for example. Moreover, GPS tracking may be used for surveillance by the government, illustrating the debate between collective security and individual privacy. It may, however, also be used in a more malicious sense within government surveillance if the state aims at identifying can religious or sexual preferences using GPS tracking, which could pose significant harm under a repressive state. It is thus important to realize that smart cities do not only pose cyber security threats, but physical security threats as well.

Potential future ethical issues in this domain include the following. To begin, relating to privacy and autonomy concerns, one may wonder what a citizen can do in a situation where his or her data is used for monetary purposes. While nowadays companies that monetise citizens' data, their services are often still avoidable (although there is often significant time and effort cost involved in doing so). In urban environments where important public services are privatized and automated (e.g., autonomous vehicles as ambulances), the options to refuse a service and avoid sharing one's are very limited. Furthermore, as the system will be increasingly connected, an issue in one part of the system may affect another part of the system as well. This may then lead to a decrease in trust in the system by citizens. Trust in the system, however, is necessary for smart cities to function properly. Finally, it has been argued that the systems used in smart cities may transform them into "panoptic cities" by the

[657] Ahmed, Kaoutar Ben, Mohammed Bouhorma, and Mohamed Ben Ahmed, "Age of big data and smart cities: privacy trade-off," *arXiv preprint arXiv:1411.0087*, 2014.

[658] Braun, Trevor, Benjamin CM Fung, Farkhund Iqbal, and Babar Shah. "Security and privacy challenges in smart cities," *Sustainable cities and society,* Vol. 39, 2018, pp. 499-507., p. 2

[659] Elmaghraby & Losavio, 2014, p. 493

[660] Braun et al., 2018

[661] AlDairi, Anwaar, and Lo'ai Tawalbeh, "Cyber security attacks on smart cities and associated mobile technologies," *Procedia Computer Science,* Vol. 109, 2017, pp. 1086-1091.

increased surveillance of data.[662] Smart cities potentially "threaten to stifle rights to privacy, confidentiality, and freedom of expression."[663]

7.1.2. Healthcare

Because of the deep power and knowledge asymmetries at its core between healthcare professionals and patients, the healthcare sector has long been guided by ethical principles, especially the Hippocratic Oath and, more recently, the principles of biomedical ethics.[664] Developments in AI raise considerable expectations in terms of improved accuracy, efficiency, cost-effectiveness, and quality they can bring to the sector.[665] As Hart puts it, "AI has the potential to revolutionize healthcare, ushering in an age of personalized, accessible, and lower-cost medicine for all."[666] The move of healthcare into the "algorithmic age", however, also brings up new ethical concerns given the profound changes in the practice of healthcare it generates.[667] This section highlights these potential ethical issues. It first identifies the more concrete and practical issues that are raised by AI in healthcare. These include potential risks to privacy and trust, gaps in accountability, threats to informed consent, discrimination, and risks of further increasing already existing health inequalities. This section concludes with remarks on the more fundamental and philosophical issues these changes raise for humanity.

The dramatic increased availability of healthcare data and the improved technological capacity to handle these data raise key privacy and confidentiality issues.[668] Furthermore, because AI technologies are primarily developed and owned by private companies, especially the 'Gang of Four' or GAFA,[669] partnerships with these companies are put in place to bring AI to healthcare. This raises significant concerns regarding the use of this sensitive personal data by these powerful commercial companies, given their rather abysmal track record on data protection, and profit-making interests.[670] If healthcare data is used by private companies against the benefits of the patients, e.g., in health insurance to

[662] Kitchin, Rob, "The real-time city? Big data and smart urbanism," *GeoJournal,* Vol. 79, No. 1, 2014, pp. 1-14.
[663] Ibid., p. 12.
[664] Beauchamp, Tom L., and Childress, James F., *Principles of Biomedical Ethics*, Oxford, Oxford University Press, 2012. We are grateful to Tally Hatzakis for reviewing this section.
[665] Wellcome Trust and Future Advocacy, "Ethical, Social, and Political Challenges of Artificial Intelligence in Health", Wellcome Trust, April 2018, pp. 12–13; Rigby, Michael J., "Ethical Dimensions of Using Artificial Intelligence in Health Care," *AMA Journal of Ethics*, Vol. 21, No. 2, 2019, p. 122; Abouelmehdi, Karim, et al., "Big Data Security and Privacy in Healthcare: A Review," *Procedia Computer Science*, Vol. 113, 2017, p. 74; Beam, Andrew L. and Kohane, Isaac S., "Big Data and Machine Learning in Health Care," *Journal of American Medical Association*, March 2018, E1–2; Microsoft, "Healthcare, Artificial Intelligence, Data and Ethics – A 2030 Vision How responsible innovation can lead to a healthier society", December 2018. https://www.digitaleurope.org/wp/wp-content/uploads/2019/02/Healthcare-AI-Data-Ethics-2030-vision.pdf
[666] Hart, Robert David, "If You're Not a White Male, Artificial Intelligence's Use in Healthcare Could Be Dangerous," *QZ*, July 10, 2017. https://qz.com/1023448/if-youre-not-a-white-male-artificial-intelligences-use-in-healthcare-could-be-dangerous/
[667] Powles, Julia, and Hodson, Hal, "Google DeepMind and Healthcare in an Age of Algorithms," *Health and Technology*, Vol. 7, 2017; Forbes Insights, "Rethinking Medical Ethics," February 2019. https://www.forbes.com/sites/insights-intelai/2019/02/11/rethinking-medical-ethics/.
[668] Abouelmehdi et al., op. cit., 2017; Char, Danton S., Shah, Nigam H., and Magnus, David, "Implementing Machine Learning in Health Care — Addressing Ethical Challenges," *New England Journal of Medicine*, Vol. 378, No. 11, March 2018, p. 3; Forbes Insights, op. cit., Feb 2019; "Amazon Alexa offering NHS health advice", *BBC News*, 10 July 2019. https://www.bbc.com/news/health-48925345
[669] GAFA are Google, Apple, Facebook, and Amazon.
[670] Powles and Hodson, op. cit., 2017; Nuffield Council on Bioethics, "Artificial Intelligence (AI) in Healthcare and Research", May 2018, p. 2.

increase premiums or reduce investment in areas of the healthcare sector that are unprofitable, then issues arise in relation to equal access to healthcare and fundamental human rights. In turn, this threatens trust in the relationships between patients and healthcare professionals and institutions. In addition, considering the highly sensitive nature of the data at stake, there is a potential issue of security related to the risk of hacking (e.g., a malevolent actor could modify the personal health data of a patient in such a way that his/her treatment may be affected in harmful ways).[671]

Trust in healthcare institutions is also potentially threatened by AI technologies. Complex AI systems, or AI systems functioning as black boxes, i.e., producing results that are sometimes hardly understandable by humans,[672] create an accountability gap that raises questions about who is to be held responsible if the system makes an error that leads to critical impact on a patient's life.[673] This complexity aspect of AI also poses a challenge for patients to give meaningful informed consent. Patients might be asked to consent to a treatment for which they were not given a proper explanation and justification as it was determined by an AI that the healthcare professional does not understand.[674] Another ethical issue relates to the increased surveillance of individual patients and the population as a whole that the use of big data analytics in healthcare implies, such as with health tracking apps.[675]

An additional set of concerns relates to potential bias and discrimination that AI may bring to healthcare practices. There is a risk that historic inequalities contained in the data that train the algorithms get entrenched. In particular, studies have shown that such healthcare data sets are largely representative of white males; hence, there is the risk that this bias in the data is reproduced by AI and that heath care needs of women and other ethnic groups be further neglected.[676] Hence, unless training databases are developed to redress these misrepresentations, the deployment of AI in the healthcare sector may further reinforce inequalities in care rather than contribute to reducing them.

A looming future concern of AI in healthcare relates to the potential deskilling of personnel.[677] As increasingly more healthcare activities are carried out by AI systems, professionals of the sector might progressively lose skills that they do not use anymore. In the future, these skills might no longer be taught in schools; this will finish to complete the replacement of humans by AI systems in these tasks. In turn, this loss of human skills is worrisome as it further exacerbates human beings' dependence on these systems.[678]

Beyond these concrete and practical issues that need to be currently addressed, more fundamental issues about the changing nature of the relationship between patients and healthcare professionals need to be considered, e.g., the impact of and connection between increasingly mechanised and

[671] Abouelmehdi et al., op. cit., 2017; Nuffield Council on Bioethics, op. cit., May 2019, p. 5.

[672] Beam and Kohane, op. cit., 2018; Sullivan, Hannah R. and Schweikart, Scott J., "Are Current Tort Liability Doctrines Adequate for Addressing Injury Caused by AI?," *AMA Journal of Ethics*, Vol. 21, No. 2 , 2019.

[673] Forbes Insights, op. cit., Feb 2019; Nuffield Council on Bioethics, op. cit., May 2018, p. 4.

[674] Wellcome Trust and Future Advocacy, op. cit., April 2018, p. 44. Nonetheless, efforts to develop explainable AI should be acknowledged.

[675] Vayena, Effy, et al., "Ethical Challenges of Big Data in Public Health", *PLoS Comput Biol.*, Vol. 11, No. 2, 2015.

[676] Hart, op. cit., 2017; Irene Y. Chen, Szolovits, Peter, and Ghassemi, Marzyeh, "Can AI Help Reduce Disparities in General Medical and Mental Health Care," *AMA Journal of Ethics*, Vol. 21, No. 2, 2019; Wellcome Trust and Future Advocacy, op. cit., April 2018, pp. 33–34.

[677] Powles and Hodson, op. cit., 2017, p. 361. On the changing nature of the healthcare profession, see also Susskind, Richard and Daniel Susskind, *The Future of Professions: How Technology Will Transform the Work of Human Experts*, Oxford University Press, Oxford, 2015. Thank you for a reviewer for pointing out this issue.

[678] Rodrigues, Rowena and Anais Resseguier, "The underdog in the AI ethical and legal debate: human autonomy", *Ethics Dialogues*, 12 June 2019. https://www.ethicsdialogues.eu/2019/06/12/the-underdog-in-the-ai-ethical-and-legal-debate-human-autonomy/

automated practices and interpersonal relationships.[679] At a deeper level, we should be aware of the changing conception of humanity that these technological transformations are bringing. An automated healthcare system that will handle digitised bodies: mechanised humanity is looming. This concern might be further exacerbated if AI moves from its "assistive role", i.e., acting as a "support tool", to progressively replacing healthcare practitioners.[680]

7.1.3. Finance and insurance

In the financial and insurance industry, AI technology is currently being deployed in a variety of ways. In finance, it is used by large institutional investors in algorithmic trading and high-frequency trading, which involves the use of complex AI systems that can perform millions of (low-margin) trades a day without human intervention.[681] In addition, AI is used in financial market analysis. Large financial institutions have invested in AI systems to assist with their investment practices and those of their clients. Such systems use big data, machine learning and natural language processing techniques to gather and analyse financial news, broker reports, social media feeds, and other sources, in order to assign ratings to potential investments. AI is also used in so-called *robo-advisors* that provide automated financial advice and in portfolio management services. These AI systems can tailor their advice and management services to the investment goals and the level of risk tolerance of a financial company's clients and can adjust in real-time fashion to changes in the market and modify portfolios accordingly.[682] Furthermore, AI is being used for underwriting purposes in the credit industry. Lenders are using machine learning techniques to analyse a large variety of variables—from purchase transactions to the manner in which a customer fills out a form—in order to develop risk models and assign scores to borrowers. Finally, in personal finance, several products have emerged that utilize AI to assist people with their personal finances, including optimization of their spending and saving practices.[683]

Insurance providers have also begun to use AI systems. They have automated some aspects of their claims processes to reduce costs, improve underwriting, improve customer experience and fight fraudulent claims.[684] Instead of relying on humans to manually comb through reports to catch fraudulent claims, insurers now often employ AI algorithms that can identify patterns in claims data and recognize attempts at fraud.

The use of AI systems in finance and insurance may raise a number of ethical issues. We will highlight some of the most important issues. First, there are issues of safety with algorithmic trading. Algorithmic trading and high-frequency trading can very occasionally induce a so-called "flash crash": a catastrophic fall in stock prices occurring within an extremely short period of time (i.e., in the order of milliseconds). [685] During such a flash crash, billions of Euros of stock value can disappear almost

[679] Char, Shah, and Magnus, op cit., 2018; Coeckelbergh, Mark, "Health Care, Capabilities, and AI Assistive Technologies," *Ethical Theory and Moral Practice*, Vol. 13, 2010.

[680] Coeckelbergh, op. cit., 2010.

[681] For a more detailed definition see http://investopedia.com/terms/a/algorithmictrading.asp

[682] Faggella, Daniel, "Machine Learning in Finanace—Present and Future Applications," *TechEmergence*, March 27, 2018. http://techemergence.com/machine-learning-in-finance-applications/

[683] Kaushik, Preetam, "Is Artifical Intelligence the way Forward for Personal Finance," *Wired.* http://wired.com/insights/2014/02/artificial-intelligence-way-forward-personal-finance/

[684] Morgan, Blake, "How Artificial Intelligence Will Impact the Insurance Industry," *Forbes,* July 25, 2017. http://forbes.com/sites/blakemorgan/2017/07/25/how-artificial-intelligence-will-impact-the-insurance-industry/#5255ab2e6531

[685] Hornigold, Thomas, "Is the Rise of AI on Wall Street for Better or Worse?," Singularity Hub, July 16, 2018. https://singularityhub.com/2018/07/16/is-the-rise-of-ai-on-wall-street-for-better-or-worse/

instantly, leaving companies and individuals with severe losses. It is posited that one of the main causes of flash crashes is the high density of very complicated, poorly understood and unpredictable automated trading agents and algorithms operating in the financial markets, which may contain design flaws and very occasionally produce errors.[686]

Second, AI systems in finance and insurance may pose security risks and could lead to instances of misuse in financial and insurance markets. It is impossible to completely guard against scenarios where such systems are corrupted by hackers or malicious designers or trainers. Since unlawful manipulation of markets through the misuse of AI systems can potentially result in enormous financial gains for a single person or group, the incentive for such malicious behaviour will be high.

Third, AI systems in finance and insurance may also raise issues of responsibility. Many AI systems in finance and insurance that are currently in use and being developed are incredibly complex, and sometimes they can be characterised as “black boxes”. The fact that many individuals are involved in the design and use of such systems, and the fact that hardly anyone has a complete understanding of the internal workings and interrelations of these systems makes it difficult to ascribe responsibility for the proper functioning of such systems, and to hold anyone accountable for any harms these systems might cause.

Finally, there is potential that the algorithms in AI systems used by lenders and insurance providers may be biased. An AI system used by an insurance company may decide to increase premiums for all individuals of particular ethnicities based on certain patterns or correlations that it found, which would result in unfair discrimination. Similarly, banks using personalized targeting could reduce an individual’s option set in life by not showing him or her crucial financial products (such as loans for education or business) due to their biased assessment of certain groups within society.

7.1.4. Defence

Use of AI in the defence sector has received much attention from an ethical standpoint over the last few years. It is essential to bring more clarity to what the use of AI in this sector actually involves. Indeed, AI is being “weaponised” in a number of different ways. This diversity is well expressed by the idea of an “AI arms race” – a very widespread expression that is used to refer to highly diverse phenomena. Peter Asaro identified seven different meanings to this expression: (1) “economic competition” between nations, (2) “proxy for technical dominance”, (3) “cyberwarfare and cybersecurity”, (4) weaponisation of AI for “social manipulation” (such as what happened in the 2016 US election), (5) “weaponizing AI for conventional warfare”, (6) “third offset strategy” (i.e., a strategy focused on “remote and autonomous platforms, big data and information processing, and information dominance”), (7) “building a super intelligence”.[687] Although all these different meanings have profound implications in the international arena and its conflictual relations, this section focuses more strictly on the use of AI in the defence sector.

Another note of caution that needs to be mentioned on the ethical issues raised by AI applications in defence is that they are generally addressed together with issues raised by robotics applications. The key element in the current debate resides in the increasing autonomy that AI technologies are bringing to the field, whether they are accompanied by a physical component or not. Considering that the most critical ethical issues are raised by robots equipped with AI, the authors of this section decided to

[686] Ibid.

[687] Asaro, Peter, “What Is an Artificial Intelligence Arms Race Anyway”, *I/S: Journal of Law and Policy for the Information Society*, Vol. 15, 2019, pp. 45-64.

explore these issues in the robotics section (subsection 7.2.3). Another reason for this choice is that, although some applications of AI in defence only consists in software elements, such as “pattern recognition systems for filtering surveillance data”[688], because they require to be addressed within the overall debate that surrounds them, they will be developed in the robotics section.

Hence, the current section strictly focuses on a particular use of AI in defence that only implies software, i.e., cyberwarfare and cybersecurity.[689] Cyberwarfare is defined as “an extension of policy by actions taken in cyberspace by state actors (or by non-state actors with significant state direction or support) that constitute a serious threat to another state's security, or an action of the same nature taken in response to a serious threat to a state's security (actual or perceived)."[690] Cyberwarfare and cybersecurity have existed without AI.[691] However, AI technologies significantly increase intensity of cyber-attacks and capacities to defend from these threats.[692] What ethical issues does this increased intensity raise in the defence sector and the society at large?

There might be a significant transformation in the conduct of war as it might shift toward the cyberspace. The potential replacement of conventional warfare by cyberwar points to the issue of the deep and increased dependence of contemporary societies on the cyber space in increasing critical ways to the point that endangering these systems threatens core functions of these societies, and hence, creates a strong vulnerability for them. Nonetheless, although this vulnerability by might be increased by AI, defence capacities to these threats are also increased thanks to AI. Hence, it is difficult to determine whether AI brings about a radical difference in this landscape with significant ethical and societal implications. According to Asaro, “it is hard to see how the incremental gains in cyber from applying AI technology could result in a dramatic strategic shift.”[693] He adds: “States likely already have the ability to wreak significant havoc, or even shut down down, each other's information infrastructures if they wanted to, without massive investments in AI.”[694] Furthermore, some experts observe that many claims on cyberwarfare are exaggerated. For instance, for Thomas Rid, “cyberwar doesn’t even exist” and “the term is being misused”.[695]

7.1.5. Law enforcement

Many ethical issues related to the use of AI across different fields of application can also be observed in its use by law-enforcement agencies (LEAs). However, considering the high stakes in law-

[688] Asaro, Peter, “Why the world needs to regulate autonomous weapons, and soon”, *Bulletin of the Atomic Scientists*, 27 April 2018. https://thebulletin.org/2018/04/why-the-world-needs-to-regulate-autonomous-weapons-and-soon/

[689] As Asaro puts it, “since AI is essentially software, ‘AI weapons’ will be cyberweapons’.” Asaro, op. cit., 2019, p. 56.

[690] Shakarian, Paulo, Jana Shakarian, and Andrew Ruel, *Introduction to cyber-warfare: a multidisciplinary approach*, Amsterda, Morgan Kaufmann Publishers, 2013.

[691] The first UN First Committee resolution on “information security” took place in 1998. UNIDIR, “The Weaponization of Increasingly Autonomous Technologies: Autonomous Weapon Systems and Cyber Operations”, UNIDIR Resources, No. 7, 2017, p. 2.

[692] UNIDIR, op. cit., 2017 and UNIDIR, “The Weaponization of Increasingly Autonomous Technologies: Artificial Intelligence”, UNIDIR Resources, No. 8, 2018.

[693] Asaro, op. cit., 2019, p. 57.

[694] Ibid., 58.

[695] Rid, Thomas, “Cyberwar – does it exist?”, *Nato Review Magazine*. https://www.nato.int/docu/review/2013/Cyber/Cyberwar-does-it-exist/EN/index.htm

enforcement activities, these ethical issues become particularly acute in this sector of application.[696] This makes an ethical analysis of the use of AI by LEAs especially critical. LEAs use AI to predict, prevent and investigate crimes.[697] Predictive policing, i.e., the analysis of "historic crime data (and sometimes other data such as social media, weather, and mortgage defaults) to predict most commonly where, but sometimes by whom or to whom, crime will take place"[698] raises numerous ethical issues as it changes profoundly the practice of policing, especially as it implies a move from "reactive policing to proactive policing."[699]

Major ethical issues at stake in the use of AI by LEAs include: (1) bias and discrimination, (2) surveillance, and (3) the accountability gap. This section focuses on these three issues and concludes with some ethical considerations on the general approach of AI by LEAs. Other affected ethical issues include: autonomy, privacy, and justice.[700]

Proponents of data mining techniques for law enforcement claim that the use of these techniques can ensure a more neutral and impartial approach to law enforcement activities because it relies on data and not human perception.[701] They claim that it removes any potential prejudices and biases, in particular those based on race. However, a number of independent studies on predictive policing tools have shown that they can actually contribute to the reproduction of historic discriminatory practices against minorities, especially black men in the US.[702] It may "lead to the over-policing of certain communities, heightening tensions, or, conversely, the under-policing of communities that may actually need law enforcement intervention but do not feel comfortable in alerting the police".[703] Selbst calls this the "disparate impacts" of AI.[704] As he puts it, this is an "artefact of the technology itself, and will likely occur even assuming good faith on the part of the police departments using it."[705] Not only does the use of AI by LEAs can reproduce existing inequalities but it might further entrench them through a self-perpetuating feedback loop. For instance, increased patrolling in a particular area leads to increased number of arrests, which, in turn, leads to increased patrolling, *etc*. As Bennett Moses and Chan note, "predictions can accordingly become self-affirming".[706] This finding should not

[696] Ferguson, Andrew, *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*, New York University Press, New York, 2017.

[697] Selbst, Andrew D. S, "Disparate Impact in Big Data Policing", *Georgia Law Review*, Vol. 52, No. 1, 2017, p. 109. Concept Paper of the 2019 OSCE Annual Police Experts Meeting *Artificial Intelligence and Law Enforcement: An Ally or an Adversary?,* 23-24 September 2019, Vienna: https://polis.osce.org/2019APEM.

[698] Bennett Moses, Lyria, and Janet Chan, "Algorithmic Prediction in Policing: Assumptions, Evaluation, and Accountability", *Policing and Society*, Vol. 28, No. 7, 2018, p. 806.

[699] Karppi, Ter, "'The Computer Said So': On the Ethics, Effectiveness, and Cultural Techniques of Predictive Policing", *Social Media + Society*, 2018, p. 1.

[700] See discussion on pre-crime in Asaro, Peter M., "AI Ethics in Predictive Policing. From Models of Threat to an Ethics of Care", *IEEE Technology and Society Magazine*, June 2019, pp. 44–46.

[701] Barocas, Solon and Andrew D. Selbst, "Big Data's Disparate Impact", *California Law Review*, Vol. 194, 2016, p. 674; Brayne, Sarah, "Big Data Surveillance: The Case of Policing", *American Sociological Review*, Vol. 82, No. 5 2017, p. 978.

[702] Civil Rights Groups, *Predictive Policing Today: A Shared Statement of Civil Rights Concerns*, 31 August 2016. https://www.aclu.org/other/statement-concern-about-predictive-policing-aclu-and-16-civil-rights-privacy-racial-justice

[703] McCarthy, Odhran James, "AI and Global Governance: Turning the Tide on Crime with Predictive Policing", Center for Policy Research, United Nation University, February 2019. https://cpr.unu.edu/ai-global-governance-turning-the-tide-on-crime-with-predictive-policing.html

[704] Selbst, op. cit., 2017, p. 109.

[705] Ibid.; See also Lum Kristian and James Johndrow, "A Statistical Framework For Fair Predictive Algorithms", 2016, p. 1. https://arxiv.org/abs/1610.08077

[706] Bennett Moses and Chan, op. cit., 2018, p. 810.

come as a surprise considering that AI systems predict the future on the basis of the past. This is the fundamental assumption at the heart of predictive policing, i.e., that "the future is like the past".[707] In spite of this assumption inherent to the technology, "predictive policing is sold in part as a 'neutral' method to counteract unconscious biases".[708] This adds another issue of ethical relevance: the obscuring of the discrimination behind the "imprimatur of impartiality on the resulting decisions".[709]

Another set of ethical issues relates to the expanded surveillance capacity facilitated by the use of AI in law enforcement. As Brayne states: "Surveillance is therefore now both *wider* and *deeper*: it includes a broader swath of people and can follow any single individual across a greater range of institutional settings."[710] This implies a radical shift in the understanding of the use of surveillance by law enforcement. While surveillance used to focus on particular suspicious individuals or groups, technology has made it possible to extend this surveillance to any individuals or groups, whether they actually are suspicious or not. Together with predictive policing, this increased surveillance capacity seriously challenges a key principle of the Western legal system – the presumption of innocence – and hence affecting the norm and value of justice. The serious impact on privacy and human autonomy of this increased capacity for surveillance by LEAs will have long term adverse impacts on individuals and society. As it has been shown, the sense of being watched generates more policed and normalised behaviours.[711] Surveillance tools are increasingly being internalised within individuals themselves, silencing any potentially dissenting voices. This certainly goes against the pluralism that the EU has as a core value.

Another issue that affects the use of AI in the law enforcement sector is common to other sectors: lack of accountability, transparency, and explainability. The issue is here the delegation of responsibility to a machine whose functioning is highly complex to the point that it sometimes appears to function as a black box. AI is a complex technology and LEAs often hardly understand its functioning. Even more problematic is that developers themselves sometimes struggle to explain the result obtained by an AI system.[712] While delegation of responsibility might not be an issue when used in sectors of application in which the stakes are rather low (e.g., the marketing sector), it becomes highly troublesome when it affects people's lives in very important ways, as it is the case in the law enforcement sector (e.g., individuals coming under suspicion and being investigated, being detained and losing their liberty and rights). Furthermore, companies often refuse to make public the content of their algorithms for trade secrecy reasons.[713] This makes the accountability requirement hard to achieve and therefore affects the capacity to maintain police forces under proper scrutiny as required in a functioning democracy. This also undermines public trust in policing.

Beyond these ethical issues inherent to the technology itself, the problem-solving approach with which AI has been implemented so far in the law enforcement sector also has deep ethical consequences that need to be highlighted. This problem-solving approach involves a technological solutionism – i.e., the view that complex social issues could be solved through a technology fix – coupled with a police

[707] Ibid.

[708] Selbst, op. cit., 2019, p. 109.

[709] Barocas and Selbst, op. cit., 2016, p. 674.

[710] Brayne, op. cit., 2017, p. 979.

[711] Greenwald, Glenn, *No Place to Hide: Edward Snowden, the NSA and the Surveillance State*, Hamis Hamilton, London, 2014.

[712] Thank you to a reviewer for help clarifying the complexity of AI systems and its implications.

[713] A US producer of risk assessments programs for policing activities refused to reveal content of its systems precisely on this ground. Barrett, Lindsey, "Reasonably Suspicious Algorithms: Predictive Policing at the United States Border", *N.Y.U. Review of Law and Social Change*, Vol. 41, 2017, p. 343.

response aimed at the symptoms rather than the root causes of security issues.[714] Finally, it is important to note that considering the high stakes in law-enforcement, new technologies in this sector need to be sufficiently tested by independent evaluators before and after deployment.[715] However, as of 2014, Uchida noted that "[t]he statistical techniques used in predictive analytics are largely untested and have not been rigorously evaluated such rigorous independent evaluation had not been conducted in the field of predictive policing."[716] For example, facial recognition is one of such AI technologies that is used by LEAs and that has raised particular concerns due to its being highly prone to error. A facial recognition system used by LEAs in China to identify jaywalkers mistakenly captured the face of a famous Chinese businesswoman printed on a bus, believing that this lady was jaywalking.[717] To conclude, as Selbst notes, "[a] great deal more study is required to measure both predictive policing's benefits and its downsides."[718] This is essential to identify the ethical issues involved in the use of AI by LEAs and to make sure appropriate measures are in place to address them.

Considering the ever-increasing interest in and use of AI by LEAs, we may fear that current concerns society is facing regarding the spread of AI in this sector are further exacerbated in the future. The risks highlighted above related to the automation of law-enforcement activities may be further exacerbated as human beings are being replaced by AI in more and more areas of the law-enforcement.

7.1.6. The legal sector

Two types of use of AI in the legal sector can be distinguished: (1) AI used to do legal research and to conduct basic legal analysis and writing tasks and (2) AI to formulate legal judgments. The former includes searching through case law or other types of documents and datasets.[719] This usage of the technology corresponds to an increased sophistication of the search capacity thanks to AI, and therefore to potentially more efficient and cost-effective access to and use of resources to support decision-making. It also includes the capacity to automate basic legal analysis and writing tasks. Proponents of the use of AI tools for such legal tasks claim that they contribute to democratising law by facilitating access to legal resources and advice.[720] However, ethical issues arise from this "computational turn" in legal practice.[721] These include risks to privacy with increasing data being

[714] Bennett Moses and Chan, op. cit., 2018, pp. 806–22; Civil Rights Groups, op. cit., 2016; Asaro, op. cit., 2019, pp. 40–53.
[715] Selbst, op. cit., 2017, p. 114; Bennett Moses and Chan, op. cit., 2018, p. 806; Civil Rights Groups, op. cit., 2016.
[716] Craig Uchida cited in Bennett Moses and Chan, op. cit., 2018, p. 815.
[717] Liao, Shannon, "Chinese Facial Recognition System Mistakes a Face on a Bus for a Jaywalker", *The Verge*, 22 November 2018. https://www.theverge.com/2018/11/22/18107885/china-facial-recognition-mistaken-jaywalker
[718] Selbst, op. cit, 2017, p. 115.
[719] The ROSS program is an example of this. See Arruda, Andrew, "An Ethical Obligation to Use Artificial Intelligence: An Examination of the Use of Artificial Intelligence in Law and the Model Rules of Professional Responsibility," *American Journal of Trial Advocacy*, Vol. 40, 2017, pp. 443–58; Nunez, Catherine, "Artificial Intelligence and Legal Ethics: Whether AI Lawyers Can Make Ethical Decisions," *Tulane Journal of Technology and Intellectual Property*, Vol. 20, 2017, pp. 189–204; Bigda, Jordan, "The Legal Profession: From Humans to Robots," *Journal of High Technology Law*, Vol. 18, 2018, pp. 396–428. Seedrs in "Six ways the legal sector is using AI right now" (13 December 2018) identifies six different aspects of this first type of use of AI in law. https://www.lawsociety.org.uk/news/stories/six-ways-the-legal-sector-is-using-ai/
[720] Arruda, Andrew, "The world's first AI legal assistant", TED Talk, November 2018. https://www.ted.com/talks/andrew_arruda_the_world_s_first_ai_legal_assistant
[721] Hildebrandt, Mireille, "The Meaning and The Mining of Legal Texts," 2010. https://www.researchgate.net/publication/41463068_The_Meaning_and_the_Mining_of_Legal_Texts

shared and mined within the legal community. It also raises potential risks of "unauthorised practice of law"[722] and mechanical interpretation of rules.[723] These evolutions due to the loss of the human element in legal practice have ethical implications as they challenge foundational aspects of the legal profession as a whole. Nonetheless, considering that this use of AI in legal practice is limited to an automatisation of basic legal tasks – not the heart of the decision-making aspect of legal practice – the ethical issues it raises are limited.

However, the use of AI in the judicial process to make high-stake legal decisions raises major ethical issues. This second type of use of AI includes in particular predictive analytic techniques that are used to provide risk-assessment to support a legal decision, a practice that is often called "predictive justice".[724] For instance, COMPAS[725] is a risk assessment software developed for the criminal justice system to assess risks of recidivism, i.e., the tendency for a convicted criminal to reoffend. The main ethical issues such systems raise are recurrent issues in the application of AI in different fields, namely (1) bias and discrimination, (2) delegation of responsibility and gap of accountability, and (3) subordination of humans to machines via relinquishment of high-impact decision making to machines.

Firstly, these predictive justice tools raise issues of bias and discrimination. Dressel and Farid note, "[p]roponents of these systems argue that big data and advanced machine learning make these analyses more accurate and less biased than humans."[726] However, a study conducted by the investigative journal *ProPublica* on COMPAS has shown that the software "was biased against blacks."[727] It was shown to overestimate black recidivism, while underestimating white recidivism.[728] In turn, this is deeply problematic for individuals and the society at large as it further heightens and intensifies discrimination, entrenches social inequalities, and covers them behind the supposed veil of impartiality of an algorithm.[729] Furthermore, Dressel and Farid have examined the accuracy of this software and concluded that "COMPAS is no more accurate or fair than predictions made by people with little or no criminal justice expertise."[730] The second set of issues raised by the use of AI to formulate risk-assessment in legal practice and sentencing concerns the delegation of responsibility it implies. Who is to be held responsible if a wrongful judgment is made based on an erroneous risk-assessment? How can one ensure accountability if the decision-making process is beyond the reach of a human or shielded from view? This issue has been identified as the "'black boxing' of the legal system", i.e., the delegation of key aspects of the decision-making process to a machine whose internal

[722] Bigda, op. cit., 2018; Simshaw, Drew, "Ethical Issues in Robo-Lawyering: The Need for Guidance on Developing and Using Artificial Intelligence in the Practice of Law," *Hastings Law Journal*, Vol. 70, 2019, pp. 173–214.

[723] Hildebrandt, op. cit., 2010.

[724] Andrew Guthrie Ferguson talks also about "predictive prosecution" which involves "identifying and targeting suspects deemed more at risk for future serious criminal activity, and then using that information to shape bail requests, charging decisions, and sentencing arguments." Ferguson, Andrew Guthrie, "Predictive Prosecution," *Wake Forest Law Review*, Vol. 51, 2016, pp. 705–44.

[725] COMPAS stands for: "Correctional Offender Management Profiling for Alternative Sanctions".

[726] Ferguson, op. cit., 2016.

[727] Angwin, Julia, Larson, Jeff, Mattu, Surya, and Kirchner, Lauren, "Machine Bias," *ProPublica*, May 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[728] Selbst, Andrew D., "Disparate Impact in Big Data Policing," *Georgia Law Review*, Vol. 52, Issue 1, 2017, p. 120. Hao, Karen, "AI is sending people to jail – and getting it wrong", *MIT Technology Review*, 21 January 2019. https://www.technologyreview.com/s/612775/algorithms-criminal-justice-ai/

[729] See European Commission for the Efficiency of Justice (CEPEJ), "European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment", adopted on 3-4 December 2018.

[730] Dressel, Julia and Farid, Hany, "The Accuracy, Fairness, and Limits of Predicting Recidivism," *Science Advances*, Vol. 4, 2018, p. 1.

functioning is opaque in addition to being undisclosed because of trade secrecy.[731] As a result, decisions made cannot be rationally explained and justified. The last ethical issue that needs to be highlighted is that of a more general subordination of human beings to machines, especially for decisions that have high-stake impacts such as a life sentence. This has major implications for human autonomy and freedom.[732]

In the future, we could imagine there will be increased automation of the practice of law, i.e., that humans may be entirely excluded.[733] This would exacerbate further the set of issues raised above and decisively facilitate the subordination of human beings to machines.

7.1.7. Public services and governance

AI is being used and significantly impacting public services[734] and governance. Will the use of smart technologies lead to smarter government and bring greater positive benefits to society? This sub-section provides an overview of two sets of potential ethical issues of the use of AI in the public sector: it first identifies issues of AI used in public services (1) and then looks at ethical risks for democratic governance (2).

Firstly (1), automation through AI technologies in public services makes it possible to conduct routine tasks more efficiently and re-allocate human resources to tasks that require more creativity; it also raises potential ethical issues.[735] It risks increasing the distance between the governed (e.g., citizens) and government and, as such, excluding further some people – particularly those with poor technological literacy. In addition, it might lead to a service that is more and more depersonalised.[736] It also poses a challenge to ensuring that it does not further exacerbate existing inequalities, but actually serves the public good.[737] Considering how essential it is for the sector to be accountable to

[731] Markou, Christopher, “Why Using AI to Sentence Criminals Is a Dangerous Idea,” *The Conversation*, May 2017. http://theconversation.com/why-using-ai-to-sentence-criminals-is-a-dangerous-idea-77734

[732] Lin et al. for instance wonder whether there are “particular moral qualms with placing robots in positions of authority […] in which humans would be expected to obey robots?” Lin, Patrick, Abney, Keith, and Bekey, George, “Robot Ethics: Mapping the Issues for a Mechanized World”, *Artificial Intelligence*, Vol. 175, 2011, p. 947.

[733] Stobbs, Nigel, Bagaric, Mirko, and Hunter, Dan, “Can Sentencing Be Enhanced by the Use of Artificial Intelligence?,” *Criminal Law Journal*, Vol. 41, Issue 5, 2017, pp. 261–77; Nunez, op. cit., 2017. At the moment, policy-makers in Europe generally do not desire such evolution for European judicial systems, as it is made clear in the “European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment”, op. cit., 2018.

[734] As defined by the European Parliament, a public service is “an economic activity of general interest defined, created and controlled by the public authorities and subject, to varying degrees, to a special legal regime, irrespective of whether it is actually carried out by a public or private body.” European Parliament, Public Undertakings and Services in the European Union. http://www.europarl.europa.eu/workingpapers/econ/w21/sum-2_en.htm

[735] UK Government, “A guide to using artificial intelligence in the public sector”, 10 June 2019. https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector; Hashimi, Ali, “AI Ethics: The Next Big Thing in Government. Anticipating the Impact of AI Ethics within the Public Sector”, World Government Summit; Deloitte, February 2019; Mehr, Hila, “Artificial Intelligence for Citizen Services and Government”, Harvard Ash Center for Democratic Governance and Innovation, August 2017; Zerilli, John and Gavaghan, Colin, “Call for Independent Watchdog to Monitor NZ Government Use of Artificial Intelligence,” *The Conversation*, 27 May 2019.

[736] Although it is also argued the opposite, i.e., that AI can actually contribute to increasing inclusion and personalization of the service. Thank you to a reviewer for raising this point. Empirical studies are needed to examine the impact of the technology on this aspect.

[737] Hashimi, A., op. cit., Feb 2019.

the public it serves, lack of transparency and explainability of the processes and decisions made by AI are problems for the use of AI in public services.[738] Government bodies and agencies collecting and analysing massive amount of data on the public also raises issues for individual freedom and privacy.[739] Another ethical issue related to the use of AI in the public sector is that it exposes users and society to great vulnerabilities (e.g., via hacking and take-down of critical public services).

The second set of ethical issues (2) concerns the challenges to justice, democracy and governance, that the deployment of AI in general brings about. We can identify three main aspects of these challenges. To begin with (2a), there is the ambiguous impact of social network platforms on democracy. On the one hand, social networks such as Facebook and Twitter have been used as platforms for civic engagement to promote democratic governance.[740] On the other hand, they have also posed a significant challenge to democracy through the mass spreading of fake news that have, in turn, contributed to an impoverishment of public political debates and a polarisation of the society through personalised political messages.[741] For instance, numerous studies have shown the role "computational propaganda" played in the 2016 US election and the Brexit referendum.[742] As a recent report commissioned by the European Parliament notes, this polarisation effect of social media on the society is the result of, both the design of these platforms (unintentional effect) and their manipulation (hence, caused intentionally by malicious actors).[743]

Another way AI constitutes a challenge to democracy and governance (2b) concerns the delegation of decision-making, responsibility, and political authority to a machine and the risks this entails for political legitimacy. As Crawford observes: "This is the more fundamental problem posed by mechanized decision-making, as it touches on the basis of political legitimacy in any liberal regime."[744] Crawford has called that "algorithmic governance", i.e., "a locus of quasi-governmental power

[738] See for instance the debate related to the use of an algorithm to offer places to students entering higher education. Graveleau, Séverin, "APB: Le gouvernement promet de se conformer aux demandes de la CNIL," *Le Monde*, 28 September 2017. https://www.lemonde.fr/campus/article/2017/09/28/mise-en-demeure-de-la-cnil-pour-changer-le-fontionnement-d-admission-post-bac_5192758_4401467.html; Hashimi, A., op. cit., Feb 2019.

[739] Gavaghan, Colin, et al., "Government Use of Artificial Intelligence in New Zealand", New Zealand Law Foundation, Wellington, pp. 46–47, 2019.

[740] For instance, they have played a key role in the Arab Spring revolutions of the early 2010s.

[741] Neudert, Lisa Maria, and Marchal, Nahema, "Polarisation and the Use of Technology in Political Campaigns and Communication," Study Panel for the Future of Science and Technology, European Parliamentary Research Service, Brussels, March 2019. http://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_STU(2019)634414

[742] Cohen, Noam, "Will California's New Bot Law Strengthen Democracy?," *The New Yorker*, 2 July 2019. https://www.newyorker.com/tech/annals-of-technology/will-californias-new-bot-law-strengthen-democracy; Howard, Philip N., Woolley, Samuel, and Calo, Ryan, "Algorithms, Bots, and Political Communication in the US 2016 Election: The Challenge of Automated Political Communication for Election Law and Administration," *Journal of Information Technology & Politics*, Vol. 15, No. 2, 2018, pp. 81–93; Johnson, Khari, "How AI Can Strengthen and Defend Democracy," *Venture Beat*, 4 July 2019. https://venturebeat.com/2019/07/04/how-ai-can-strengthen-and-defend-democracy/. Gallagher, J., et al., "Junk News and Bots during the 2017 UK General Election: What Are UK Voters Sharing Over Twitter?," Data Memo, COMPROP-OII, May 2017. https://comprop.oii.ox.ac.uk/research/working-papers/junk-news-and-bots-during-the-2017-uk-general-election/; Helbing, Dirk, et al., "Will Democracy Survive Big Data and Artificial Intelligence?," *Scientific American*, 25 February 2017. https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/

[743] Neudert and Marchal, op.cit., p. 3.

[744] Crawford, Matthew B., "Algorithmic Governance and Political Legitimacy," *American Affairs*, Vol. III, No. 2, 2019.

untouched by either the democratic process or by those hard-won procedural liberties that are meant to secure us against abuses by the (actual, elected) government".[745]

Finally, the third ethical risk (2c) of the deployment of AI for democracy and governance is that this technology provides extremely powerful tools to impose a totalitarian regime. AI renders surveillance possible with both an amplitude and a precision never thought possible before.[746] This risk is already well-illustrated by China's use of digital technology that demonstrates a move toward widespread state surveillance and social control.[747] Although the EU might feel insulated from such totalitarianism, democracy should never be taken for granted, and tools that would be extremely useful to a dictatorship or totalitarian regime pose a high risk to future democracy and individual rights and civil liberties.

The current concerns highlighted above might be further exacerbated in the future as AI technology is integrated in increasingly more sectors of public services and in the society at large to the point that the heart of governance shifts from human beings to automated systems.

7.1.8. Retail & marketing

Ethical issues with AI applications in retail and marketing fall into five main categories: issues of privacy, issues of autonomy, issues of discrimination, issues of sociality, and issues to do with the inaccuracies produced by, and overconfidence in, the AI systems used for retail and marketing purposes. First, big-data-driven AI systems for personalised advertising may lead to issues of privacy and data protection. Inherently, the processing of large amounts of sensitive personal data (e.g., information from tracking cookies recording personal browsing history) brings with it substantial privacy and data protection risks. Further, AI-based profiling in marketing permits far-reaching identification of consumers' preferences and personalities. Consumers may sometimes even not have been aware they had certain preferences before a personalised advertisement targeted those preferences. This might make them prone to manipulation. The use of data across social contexts in AI-based advertising may also be problematic in many instances (e.g., browsing data from a pornographic website influencing personalized advertising on Amazon's website). An example illustrating the potential for privacy harms is the story of a US teenager whose web browsing behaviour seemed to indicate she was pregnant, and whose parents therefore got sent mail advertisements for maternity clothing and nursery furniture, which promptly revealed the teenager's actual pregnancy.[748]

Second, AI-driven micro-targeting and nudging practices in marketing and retail may have a negative effect on consumers' autonomy. Through psychological reductionism on the basis of recorded consumer behaviour, these practices may detract from autonomy in three ways. First, such practices may ignore so-called *meta-preferences* (i.e., preferences about one's preferences) that are generally inaccessible to the algorithms. Through targeted advertisements based on their past behaviour, consumers may be have to fight harder against their own bad impulses to make better choices for

[745] Ibid. See also the concept developed of "algorithmic governmentality" developed by Antoinette Rouvroy; Rouvroy, Antoinette, and Berns, Thomas, "Algorithmic Governmentality and Prospects of Emancipation," *Réseaux*, Vol. 177, No. 1, 2013.

[746] Brayne, Sarah, "Big Data Surveillance: The Case of Policing," *American Sociological Review*, Vol 82, No. 5, 2017.

[747] Mitchell, Anna and Diamond, Larry, "China's Surveillance State Should Scare Everyone," *The Atlantic*, 2 February 2018. https://www.theatlantic.com/international/archive/2018/02/china-surveillance/552203/

[748] Lubin, Gus, "The Incredible Story Of How Target Exposed A Teen Girl's Pregnancy," *Business Insider*, February 12, 2012, https://www.businessinsider.com/the-incredible-story-of-how-target-exposed-a-teen-girls-pregnancy-2012-2?international=true&r=US&IR=T.

themselves (e.g., a person giving in to her urge to smoke while she would have actually liked to quit).[749] Second, such practices may also deprive consumers of opportunities for introspection regarding their preferences, by offering them a personalised set of attractive options that they can easily choose from, thus eliminating the need for careful deliberation and weighing of options.[750] Third, these practices may lead to consumers getting confined in "information bubbles" that are difficult to escape out of. The focus on making highly personised recommendations based on past choices could reinforce a consumer's present patterns of consumption and keep those patterns from evolving over time, or at least reduce the chance of radical change occurring in the person's tastes.[751]

Third, AI-based personalised marketing and retail practices and general AI-based automation may lead to unfair discrimination for consumers. AI technology may enable highly personalised pricing practices based on people's past shopping behaviour and inferences on their financial means, which may not always be fair. Also, there is a risk that factors we might not want AI-based systems to base their decisions on will still be taken into account, such as race or ethnicity. Further, in retail, there may be a higher incentive for AI-based automation at the more price-competitive lower end of the market than at the higher end,[752] which could potentially result in consumers on a budget being largely relegated to shopping online, whereas well-to-do individuals can afford to buy at high-end physical stores where they are served by human sales clerks.

Fourth, and related to the previous point, AI technology in online retail (e.g., shopping by intelligent assistants, AI-based order fulfilment, virtual helpdesk agents) may have a negative effect on people's sociality and sense of community, and therefore their well-being.[753] For many consumers, shopping in physical stores is not merely an activity focused on procuring the goods they need, but is also a social activity in that it allows them to interact socially and build social relationships, and in that it contributes to a sense of community. AI technology that enables higher cost-efficiencies and more convenience in online retail may further accelerate the disappearing of physical stores, especially on the lower end of the market, thus diminishing traditional opportunities for people to interact socially with one another.

Fifth and finally, there may be a potential for harms caused by the making of inaccurate inferences by the AI systems used in retail and marketing, and overconfidence in the accuracy of the inferences by these systems. Sometimes, the preferences and personality characteristics inferred from one's online behaviour can be wildly different from one's actual preferences and personality. Many situations can be imagined where this can lead to harms. An example can be a grieving mother of a stillborn child who continued to be served motherhood advertisements on the basis of her past search history.[754]

7.1.9. Media & entertainment

The media and entertainment industry consist of companies whose business model centres around the communication of information, art and entertainment to a large audience. It includes publishing

[749] André, Quentin, et al., "Consumer Choice and Autonomy in the Age of Artificial Intelligence and Big Data," *Customer Needs and Solutions*, Vol. 5, p. 28–37.
[750] Ibid.
[751] Ibid.
[752] The overheads of staffing and operating physical stores can be removed through automation and home delivery.
[753] SIENNA London expert workshop on the identification of future social and ethical issues in AI&R, January 2019.
[754] Feiner, Lauren, "A woman shared her tragic story of how social media kept targeting her with baby ads after she had a stillbirth," *CNBC*, December 12, 2018, https://www.cnbc.com/2018/12/12/woman-calls-out-tech-companies-for-serving-baby-ads-after-stillbirth.html

(including books, magazines, newspapers, online news, blogs, etc.), film, music, video games, broadcast radio and audio productions, graphics, streaming and interactive media, social media, and theatre and art. Also included are information service companies that offer search engines, online databases, etc. The media and entertainment industry is a large and diverse industry, and AI is being used in it in many different ways, impacting all media and entertainment industries, and all parts of the media value chain, from early planning to content creation to distribution. We will limit our scope to those applications of AI that raise the most significant ethical issues.

In what follows, we will focus on three types of media and entertainment in which the application of AI has raised particular ethical issues. These are (1) audio and visual media; (2) news media, and (3) social media.

Audio and visual media include film, video, audio, video games, music, and graphics. AI is currently acquiring a major place in content creation and production.[755] In film, it already has a major role in animation and the creation of special effects, and is increasingly being used to produce photorealistic simulations of human beings, including lead actors. It is also being used in the navigation and retrieval of media content in large databases. It is similarly being used in graphic arts to produce images, including photorealistic images of scenes and people. In music, it is being used to create new compositions based on examples by retaining elements of their underlying structure, to make style transfers, and to de-mix music pieces to retain particular sounds and loop, to enhance audio quality and support the mixing process. In video games, it is being used to automate the process of creating game environments, to create characters and environments that respond intelligently, to optimize games for different human skill levels, and to personalize the game experience based on knowledge about the player's skills, preferences and mental states.

In all these industries, AI is also being used to reformat and repurpose content. In the movie industry, for example, AI can automate the process of making adaptations of movies for different international markets, even to the extent that the facial expressions of animations are changed to better synchronize with foreign voices. It is also being used to predict demand in different markets. And AI is being used to better target consumers to find content that matches their interests, amongst others by recommender systems.

At its best, AI helps content creators and studios to harness their creativity and to outsource mundane and repetitive tasks to AI, and it helps consumers find and enjoy the media they are most interested in, and possibly to give them a personalized experience. However, AI can also do harm in this industry. We will now discuss several ethical issues.

Most importantly, if AI is pushed too far, and takes over the creative process, rather than aiding graphic artists, musicians, actors and directors, and other content creators in their creativity, then content creation can become a formulaic process and miss the creativity, spontaneity and humanity that (arguably) only human beings can bring. We will discuss two cases of AI going too far in this sense. A first is the possible future replacement of human actors by digital ones. Animation and CGI are already limiting some actor roles to voice acting and acting for motion capture, and future AI may eliminate

[755] NEM, "Artificial Intelligence in the Media and Creative Industries," Position paper, July 2018, https://nem-initiative.org/wp-content/uploads/2018/10/nem-positionpaper-aiinceativeindustry.pdf

the need for human actors altogether.[756] By giving up on human actors, one would however risk losing the authenticity and humanity that only human actors can bring.

Second, AI algorithms are now being developed that could predict how much money a movie would make, by comparison with past movies and recent trends, using machine learning.[757] Such algorithms would help investors to maximize their profits, but would also make the movie industry even more conservative than it already is, by only allowing movies to be made that are similar to past successful movies. The use of personal profiles and recommender systems for consumers could similarly limit creativity and diversity, by only exposing consumers to content that is similar to what they have liked in the past, and thereby confining them to an entertainment "filter bubble".

Now let us discuss ethical issues in relation to news media. AI is currently transforming news media. It is being used to collect data, to analyse large data bases, to track down breaking news and trends, to produce stories from data ("automated journalism"), to generate scripts for video production, to support data visualisation, to do automated fact checking, to eliminate reporting bias, to recognize and eliminate fake news, to engage with audiences using newsbots, and to personalize news feeds and even news content. Potentially, these applications of AI can augment journalists and newsrooms, as well as news consumers. However, several of them also raise ethical issues.

One concern is that systems for automated journalism, involving the collection and analysis of data and the production of stories, do not necessarily abide to the values of journalism.[758] Algorithms may be biased, may mislead, may human make unwarranted inferences (though correlations) and claims and may violate the rights of data subjects and other parties.[759] They may not uphold journalistic standards of impartiality, accuracy, independence, humanity and accountability. Transparency, an important value in journalism, is also at risk. Machine learning systems are typically not transparent, and may not reveal how they collect and analyse data. Systems that are not transparent may not have their inner workings exposed because they contain proprietary software.

Another concern is that AI could end up impoverishing journalism by replacing journalists and making it difficult for smaller newsrooms to cope. The latter danger results from the fact that smaller, local newsroom currently cannot afford expensive AI systems. Larger newsrooms can therefore gain a competitive advantage and use their AI technology to generate local news that competes with the local news generated by small local newsrooms.[760]

A third issue concerns the dangers of hyper-personalization that could come from the application of AI and machine learning to generate personalized newsfeeds and even news stories for users. As discussed in the section on autonomy and liberty, personalisation of news and information feeds runs

[756] Kemp, Luke, "In the Age of Deepfakes, Could Virtual Actors Put Humans out of Business?," *The Guardian*, Guardian News and Media, July 8, 2019. https://www.theguardian.com/film/2019/jul/03/in-the-age-of-deepfakes-could-virtual-actors-put-humans-out-of-business

[757] Vincent, James, "Hollywood Is Quietly Using AI to Help Decide Which Movies to Make," *The Verge*, May 28, 2019. https://www.theverge.com/2019/5/28/18637135/hollywood-ai-film-decision-script-analysis-data-machine-learning

[758] OSF Journalism, "Artificial intelligence demands genuine journalism," *Medium*, October 31, 2018. https://medium.com/innovation-in-journalism/artificial-intelligence-demands-genuine-journalism-8519c4e0fc86

[759] To be fair, human journalists may at times also be guilty of these things.

[760] Dossett, Julian, "Artificial Intelligence: Raising New Ethics Questions in Media and Journalism," *PR Newswire for Journalists*, May 9, 2018. https://mediablog.prnewswire.com/2018/05/09/artificial-intelligence-ethics-questions/

the risk of enclosing people in filter bubbles in which their horizon is limited and their opinions and prejudices are confirmed.

A final issue concerns the application of AI to generate and distribute fake news. AI programs exist that are capable of generating very convincing news stories, that are even more believable than stories written by humans.[761] The distribution of fake news stories, whether propagated by adversary foreign powers, individuals and groups that seek to monetize content, or by others, causes social harm by instilling false beliefs, corrupting democratic processes, and undermining trust in the news media. In this light, the emergence of deepfakes brings particular worries. Deepfakes are images and videos that involve the combination of multiple images or videos through machine learning to produce fake images and videos that are very difficult to identify as fake. Not only can deepfakes spread fake news and false beliefs, they also undermine confidence in any recently produced image or film, as it could also be a deepfake.[762]

Finally, let us turn to ethical issues in social media. AI already has a central role in social media. It is being used to index, search and use social media content. Both text, images and videos on social media are being analysed and mined for information using AI. Since most social media companies have a business model that is based on targeted advertising, AI is being used for profiling and targeted advertising using highly distributed recommender systems. AI is also being used for monitoring and removal of content that violates company policy, and for improvement of services. It is also being used for various types of opinion mining and trend detection based on social media data.

Social media has been the subject of many recent scandals, most of which concern the use of personal data of social media users. Social media contain very rich personal data, giving insight into people's traits, habits, behaviours, preferences, social relations, and personal histories. It is this very data that are being exploited and monetized in the business model of most social media companies, for the purposes of targeted advertising and messaging. This microtargeting risks violating the user's privacy, many have argued.[763] When it is being used to manipulate public opinion to promote political ideals, as has been claimed to happen in the Facebook/Cambridge Analytica scandal, it may even undermine democracy (Margetts, 2019).

AI algorithms can also contribute to the generation of echo chambers on social media: online communication spaces in which like-minded beliefs and ideas are reinforced through repetition in a closed system that does not allow for alternative viewpoints and can reinforce extreme views. Much evidence has been presented that such echo chambers exist on social media (Williams et al., 2015; Quattrociocchi, Scala and Sunstein, 2016). There is evidence that the AI-driven recommender algorithms in social media stimulate the formation and persistence of echo chambers, as well as corresponding filter bubbles (Jiang et al., 2019; Sasahara et al., 2019).

Social media censorship, finally, also raises significant ethical issues. Such censorship takes place by social media companies and by governments. All social media companies have policies for banning objectionable content, and often employ AI algorithms for detecting and eliminating such content,

[761] Robitzski, Dan, "New AI Generates Horrifyingly Plausible Fake News," *Futurism*, May 30, 2019. https://futurism.com/ai-generates-fake-news.

[762] Hall, Holly Kathleen, "Deepfake Videos: When Seeing Isn't Believing," *Catholic University Journal of Law and Technology,* Vol. 27, No. 1, 2018, pp. 51-76. https://scholarship.law.edu/jlt/vol27/iss1/4.

[763] Wilson, Dennis G, "The Ethics of Automated Behavioral Microtargeting," *AI Matters,* Vol. 3, No. 3, 2017, pp. 56-64.; Jacobson, Jenna, Anatoliy Gruzd, and Ángel Hernández-García, "Social media marketing: Who is watching the watchers?," *Journal of Retailing and Consumer Services*, available online March 20, 2019, in press. https://doi.org/10.1016/j.jretconser.2019.03.001.

next to human intervention. Governments can similarly use automated systems to censor social media posts. This happens especially in countries with authoritarian governments, like China and Saudi Arabia. In liberal democratic societies, content that is not allowed typically includes content that promotes or publicizes criminal acts, that glorifies violence and enjoys suffering, that displays nudity or sexual activity, that is cruel or insensitive to the misfortune of others, that violates intellectual property rights, that promotes false news, and that contains hate speech. Most controversial of these are the curtailment of hate speech and false news. Opponents of such censorship hold that social media companies should not be in the business of deciding whether or not speech is hate speech, or whether news is fake news, and that these companies and their AI algorithms can harbour political biases because of which such censorship is not even-handed, and that hate speech should not be censored to begin with, as it should be seen as protected free speech (Heins, 2014; Strossen, 2018). The censorship of social media content is likely to remain a topic of moral controversy in the future.

7.1.10. Smart home

Smart Home technologies are applications of embedded intelligence and Internet of Things technologies. Smart home technologies are used inside the house by residents to increase efficiency of their home and to improve security. Examples include Amazon Echo and Nest Cam Indoor security. While there are potential benefits from the use of SH technologies, such as increased feelings of safety for the residents or improvements to assisted living, the technologies also raise ethical concerns. As elaborated on in the earlier subsection on embedded intelligence and Internet of Things, concerns relating to privacy and freedom and autonomy are some of the most pressing issues. Indeed, smart home technologies illustrate these concerns. As Internet-of-Things (IoT) devices are becoming more diverse and accessible, people's lives are increasingly recorded and documented. It is therefore worth reflecting on the ethical concerns around technologies related to smart homes.

The reason smart home technologies are so privacy sensitive is not only due to their ability to communicate with one another, but it is the 'always-on' mode that most of such technologies have adapted. An always-on mode allows the devices to constantly monitor the behaviour of the resident, his or her needs, in order to reach the highest levels of operational performance possible. For example, speech recognition systems such as Amazon Echo remain in an always-on mode to allow themselves to receive the trigger word ("Hello Alexa"). While they are not actively recording in this always-on mode, errors do occur and they might start recording when the trigger word in fact had not been spoken by the user, a situation that raises privacy concerns.[764,765] Indeed, research has shown that intelligent virtual assistants (IVAs) such as Amazon Echo and Google Home are not all that trustworthy.[766] Chung et al. have analysed four ways in which IVA-enabled devices may constitute security and privacy threats. Firstly, personal information may be wiretapped. Most IVA-enabled devices do not use "encrypted connections to check network connectivity" when connecting with the cloud, which then allows for other SH devices to be detected.[767] Secondly, IVA-enabled devices may be

[764] See also the subsection on Natural Language Processing

[765] E.g., Henderson, Peter, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau, "Ethical challenges in data-driven dialogue systems," In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 123-129, ACM, 2018.; Lau, Josephine, Benjamin Zimmerman, and Florian Schaub, "Alexa, are you listening?: Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers," *Proceedings of the ACM on Human-Computer Interaction,* Vol 2, No. CSCW, 2018, p. 102.

[766] Chung, Hyunji, Michaela Iorga, Jeffrey Voas, and Sangjin Lee, "Alexa, can I trust you?," *Computer,* Vol 50, no. 9, 2017, pp. 100-104.

[767] Ibid., p. 102

compromised due to the "always-on" mode of the devices. This allows hackers to "monitor all voices and sounds within the device's range in real time."[768] Thirdly, the hacking of SH devices opens up the possibility for malicious voice commands, which may result in for example theft. Lastly, as mentioned earlier, the device may record conversations while the user is unaware of this. Either the device incorrectly hears the trigger word and starts recording, or the device records independent of the trigger word, as may be the case with Google Home, for example.[769] In the case of Google Home, users do have the choice to not share their recordings with Google. However, this severely impacts the efficiency of the system, limiting its use greatly. In this case, convenience and privacy are seemingly at odd with each other.

A few other ethical concerns are also worth mentioning. For example, speech recognition devices may exhibit biases. When the training data lack voice data of women and minorities, these groups may become more difficult to understand for these systems.[770] In addition, voice assistants tend to display a gender bias in the sense that most systems make use of a female voice, thus possibly further reinforcing the stereotype of a women as assistants.

Smart home technologies can be used to provide social support. The technologies may be used for healthcare assistance in order to "support people to have a better quality of life and to ensure elderly to live comfortably and independently."[771] Linked through the Internet-of-Things by using sensors, smart calendars, and so forth, Smart Home technologies may "control the environment on behalf of the residents, predict their actions and track their health condition."[772] Companionship is an important component of friendship, which may contribute to someone's happiness.[773] Research shows that smart home devices may positively affect the perception of companionship by providing social support.[774] This contradicts the common worry that smart technologies may increase a feeling of isolation in the elderly people, for example. Sensors are used to monitor the technologies and the residents and so are able to formulate a perception of the environment (e.g. is the door open? Is the resident in bed?).

A current problem is that most smart technologies are designed without explicitly considering ethical values and concerns.[775] Ienca et al. have shown in an extensive literature review that maintaining one's autonomy and independence is considered to be of great value for people with dementia.[776] Maximizing a user's autonomy implies taking into account his or her needs. Rather than a top-down design, responsiveness to individual needs is thus of great importance. Other ethical concerns include justice (in that some individuals may be able to afford a particular smart home technology, whereas others might not), and the potential for social isolation by making it easier for users to withdraw

[768] Ibid., p. 102

[769] See https://www.theguardian.com/technology/2019/jul/11/google-home-assistant-listen-recordings-users-privacy

[770] See also the subsection on Natural Language Processing

[771] Amiribesheli, Mohsen, Asma Benmansour, and Abdelhamid Bouchachia, "A review of smart homes in healthcare," *Journal of Ambient Intelligence and Humanized Computing,* Vol. 6, No. 4, 2015, pp. 495-517., p. 495

[772] Ibid., p. 496

[773] Lee, Byounggwan, Ohkyun Kwon, Inseong Lee, and Jinwoo Kim, "Companionship with smart home devices: The impact of social connectedness and interaction types on perceived social support and companionship in smart homes," *Computers in Human Behavior,* Vol 75, 2017, pp. 922-934.

[774] Ibid., 2017

[775] Ienca, Marcello, Tenzin Wangmo, Fabrice Jotterand, Reto W. Kressig, and Bernice Elger, "Ethical design of intelligent assistive technologies for dementia: a descriptive review," *Science and engineering ethics,* Vol 24, No. 4, 2018, pp. 1035-1055.

[776] Ibid.

themselves from society. The latter might negatively affect users' quality of life by reducing their physical or emotional wellbeing.[777]

7.1.11. Education & science

AI has a large number of applications in education and in scientific research. We will discuss key applications and the ethical issues that they raise. Let us first discuss some of the main ethical issues in education. AI may impact the learning and instruction process in several profound ways, enhancing and supplementing teachers and administrators, and supporting students in their learning activities.[778] The first way is by enabling smart content: interactive digital course materials that replace or supplement textbooks. This type of content is customizable and personalisable, and can break down and explain textbook content through flashcards, chapter summaries, practice exercises and tests, real-time feedback and comprehensive assessments. AI may also be used to optimize study materials through machine learning.

Secondly, AI is being used to develop intelligent tutoring systems. These are systems that actively tutor students by explaining basic concepts, theories and methods, taking into account aspects like the learning history, cognitive style and preferences of the student. Currently, these tutoring systems are not expected to be able to replace teachers, since they are not capable of attaining the advanced expertise and pedagogical skills of teachers. However, they could be a good supplement for some students. Thirdly, smart learning environments that use AI, 3-D gaming, computer animation and augmented reality can create new learning environments that involve realistic virtual characters and social interactions and that may offer new instructional and learning opportunities.

Fourthly, AI can support monitoring practices in education: the continuous assessment and evaluation of the quality and effectiveness of instruction and the progress of students. AI systems are good at monitoring, if fed the right data. They can provide helpful feedback to teachers, administrators and students. AI systems are already in use that monitor student progress, by logging online behaviour or assessing overall progress, and to alert professors when there might be an issue with student performance. Other systems can identify and correct weaknesses in courses based on assessments of student performance and propose remedies. Fifthly and finally, AI is able to automate student assessment, including grading of tests and student admissions. AI can grade and provide feedback on tests and essays, and automate the classification and processing of paperwork in admission processes and make recommendations for admissions.

While AI has clear potential benefits for education, there are pitfalls as well. First of all, there is the risk that AI will be used as a cheaper alternative to teaching by actual teachers, and provide inferior quality. Since AI systems have not shown to be capable of attaining the expertise level of teachers – both in subject matter and in didactic and pedagogical skills, they are still inferior to real teachers. Secondly, uncritical adoption of AI in education can lead to unfair treatment of students and pupils. The reliance on inferior AI systems for assessment could lead to unfair practices in testing and admissions. Systems

[777] Ibid.

[778] Uskov, Vladimir L., Robert J. Howlett, and Lakhmi C. Jain, (eds.), *Smart education and smart e-learning*. Vol. 41, Springer, 2015.; Utermohlen, "4 Ways AI Is Changing the Education Industry," *Medium*, Towards Data Science, April 12, 2018. https://towardsdatascience.com/4-ways-ai-is-changing-the-education-industry-b473c5d2c706

may also contain biases and prejudices that are to the disadvantage of some students.[779] Decisions and recommendations by AI systems, for example in testing or admissions, may moreover be difficult to challenge if the technology is not transparent. Another issue is that extensive monitoring of students and pupils (and teachers) by AI systems can potentially undermine their privacy.[780]

Let us now turn to ethical issues in in the field of science. The main application of AI in science today is in data mining. In many fields, including natural sciences, engineering sciences, medical and life sciences, and social sciences, advances in research increasingly depend on the creation and mining of large data sets. The use of AI is radically changing scientific investigation by facilitating the production and analysis of large data sets. Such analysis is used to uncover deep patterns and correlations in data, and to develop predictive insights.

AI can be applied at almost every step in the research process. AI is used in observation and data collection through the use of smart sensors and AI programs that manage the data collection process. Using deep learning and other methods, collected data can be cleansed and analysed to uncover deep patterns. While AI programs will not for the foreseeable future be able to formulate broad research questions, which remains the prerogative of researchers, they can raise insights into how to further specify a research question or hypothesis by uncovering relationships in data that suggest promising ways of (re)formulating the research question. AI programs may even generate large numbers of hypotheses and test them against data, thus suggesting more promising or valid hypotheses for scientists to consider.[781]

In general, AI can increase the efficiency of research by automating the more routinised, labour-intensive research activities, like literature search, data collection, clustering and hypothesis testing. However, more interpretive, reflective and creative tasks still remain the prerogative of the researchers, such as formulating research questions, reading and synthesising prior literature, flagging gaps and inconsistencies and omissions, developing new concepts and theories, and writing up publications.

The use of AI in scientific research also brings several ethical issues to consider. First of all, transparency can become an issue. If data analysis and hypothesis testing are the result of machine learning, then scientists may not be able to explain how they have arrived at their conclusions, and this undermines the transparency of scientific inquiry, as well as the ability of third parties to challenge results. Secondly, algorithms may contain biases and prejudices, and may therefore lead to biased outcomes without this being known. This could threaten the quality of science overall, but would be especially worrisome in the humanities and social sciences, as it could stigmatize and discriminate against social groups. Thirdly, the vast amounts of data generated in contemporary science raise issues of access to data and data ownership, especially in the context of private enterprise. Fourthly, the collection and analysis of personal data also raises issues of privacy and informed consent, especially if new uses are made of personal data that data subjects have not explicitly consented to. Fifthly, the use of AI in science brings new challenges for research integrity and social responsibility for scientists, since they have to maintain these virtues as they delegate significant aspects of their activities to machines that

[779] Popenici, Stefan A., and Sharon Kerr, "Exploring the impact of artificial intelligence on teaching and learning in higher education," *Research and Practice in Technology Enhanced Learning,* Vol. 12, No. 1, 2017, p. 22. DOI 10.1186/s41039-017-0062-8.; Johnson, Jeffrey Alan, "The ethics of big data in higher education," *International Review of Information Ethics,* Vol. 21, No. 21, 2014, pp. 3-10. http://www.i-r-i-e.net/inhalt/021/IRIE-021-Johnson.pdf

[780] Johnson, 2014

[781] Kitano, Hiroaki, "Artificial intelligence to win the nobel prize and beyond: Creating the engine for scientific discovery," *AI magazine* Vol. 37, No. 1, 2016, pp. 39-49. DOI: https://doi.org/10.1609/aimag.v37i1.2642.

they do not fully control and may not fully understand. Finally, the use of AI and large data sets in science might bring with it the risk of science becoming “theory-free”, i.e., correlations identified by AI are replacing causal theories, which may pose a number of risks (e.g., in relation to the validity of scientific claims).[782] (For more extensive discussion of ethical issues concerning the application of AI in science, see section 7 of the 2019 SHERPA report on ethical tensions and social impacts.[783])

7.1.12. Manufacturing

When deployed in manufacturing, AI generally comes together with a physical component; hence, most of the ethical issues it raises are identified in the robotics section of this report. However, as the SIENNA State-of-the-art Review report (D4.1) highlights, AI as software is also present in the manufacturing sector, including for predictive maintenance of industrial equipment, for automated quality control, and for demand-driven production.[784] It enables the automation of tasks and activities that, until recently, remained in the hands of human beings as they required abilities of attention and flexibility that automated systems could not yet exhibit.[785] The prospects of AI in manufacturing are promising, especially to improve product quality, performance of industrial systems, and ensure appropriate production levels (i.e., avoiding over or underproduction). However, they also raise ethical questions.

A key ethical issue, with significant social and economic underpinnings, is the threat to employment. Like previous industrial revolutions, the one brought about by AI is deeply impacting the labour market.[786] There are fears that it might “exacerbate societal inequalities by reducing employment and wages—especially for the working and lower middle classes.”[787] There are conflicting views as to whether automation will necessarily be accompanied by loss of jobs.[788] However, a number of studies shows that the losers of this transformation might primarily be those with middle-skills occupations.[789] Experts have identified this trend as the “polarisation” of the job market, with the rise of both “lousy” and “lovely” jobs.[790] AI systems now able to automate “‘routine’ white-collar jobs” significantly contribute to this trend of “hollowing-out of middle-income routine jobs.”[791]

Furthermore, working conditions and well-being at work might be significantly affected in ways that pose ethical challenges.[792] For instance, employees might find themselves subordinated to machines

[782] Anderson, Chris, “the end of theory: the data deluge makes the scientific method obsolete,” WIRED, June 23, 2008. https://www.wired.com/2008/06/pb-theory/

[783] Ryan, Mark, Philip Brey, Kevin Macnish, Tally Hatzakis, Owen King, Jonne Maas, Ruben Haasjes, Ana Fernandez, Sebastiano Martorana, Isaac Oluoch, Selen Eren, and Roxanne Van der Puil, *Report on Ethical Tensions and Social Impacts.* SHERPA Project, 2019, https://doi.org/10.21253/DMU.8397134.

[784] Jansen, Philip, et al., “State-of-the-art Review”, AI & Robotics, SIENNA Deliverable (D4.1), March 2018.

[785] Acemoglu, Daron, and Restrepo, Pascual, “Artificial Intelligence, Automation and Work”, National Bureau of Economic Research (NBER), January 2018, p. 4.

[786] Petropoulos, Georgios, “The Impact of Artificial Intelligence on Employment”, in Neufeind, Max, O’Reilly, Jacqueline, Ranft, Florian, *Work in the digital age*, Rowman and Littlefield, London, 2018, pp. 119-132.

[787] Wright Scott A., and Schultz, Ainslie E., “The Rising Tide of Artificial Intelligence and Business Automation: Developing an Ethical Framework,” *Business Horizons*, Vol. 61, 2018, p. 824.

[788] Acemoglu and Restrepo, op. cit., Jan 2019, p. 4.

[789] Frey, Carl Benedikt, and Osborne, Michael A., “The Future of Employment: How Susceptible Are Jobs to Computerisation?,” unpublished, September 2013, p. 3.

[790] Frey and Osborne, op. cit., Sept 2013; Reuters, “The Rise of Lousy and Lovely Jobs”, 13 April 2012. https://www.reuters.com/article/idUS380409786120120412

[791] Frey and Osborne, op. cit., Sept 2013, p. 3; Reuters, op. cit., April 2012.

[792] Trentesaux, Damien and Raphael Rault, “Designing Ethical Cyber-Physical Industrial Systems,” *IFAC PapersOnLine*, Vol. 50, No. 1, 2017.

and subjected to a more advanced system of surveillance and monitoring – to a degree never thought possible before.[793] In turn, this may profoundly affect their sense of autonomy and dignity. In addition, in view of the rapid pace at which AI develops and enters different sectors of the society, including manufacturing, employees might need to rapidly and regularly switch (often without choice) to new occupations with very less time to re-train and/or re-skill; that might further increase their stress.[794] Considering that work is one of the key sources of the sense of self-worth and well-being in contemporary society, automation and potential disruption in employment caused by AI may have deep effects on individuals.[795] AI in manufacturing also poses the question of responsibility: who is responsible or could be held responsible in case an error is caused by an AI system?[796] Would it be the designer, the manufacturer, the supplier, the system integrator, the user, or the owner?

Another area of ethical concern relates to the increased standardisation in industrial products that the use of AI in the manufacturing sector might generate. AI in this sector further intensifies the process of standardisation that the development of the factory systems witnessed from the second half of the 18th century that replaced handicraft.[797] This comes together with "deskilling", i.e., the disappearance of handicraft skills as means of production shifted to the factory.[798] Standardisation and loss of skills have ethical underpinnings as they lead to a loss of diversity and to a continuously increased homogenisation of the society.

In the future, we might fear the issues highlighted above to be further exacerbated. In particular, we may be concerned by a radicalised subordination by the automated systems thanks to the choice of AI over the human. Furthermore, this had profound and worrisome implications in terms of the objectification of human beings, i.e., the mechanisation of human activities and behaviour as they are led to increasingly interact with machines.

7.1.13. Agriculture

The digital revolution is also impacting the agricultural sector. Data-driven farming offers the potential benefits of an agriculture that is led by more precise, accurate, and timely analysis and therefore potentially more effective and cost-efficient; its proponents hence claim that it may significantly improve productivity.[799] However, 'smart farming' also raises a number of ethical issues that this section aims to identify. This section highlights two sets of ethical issues the use of AI in agriculture raises. First, it explores issues related to the power asymmetry at stake between farmers and powerful agribusinesses (i.e., companies in the business of agricultural production such as John Deere or Monsanto) and how AI risks further exacerbating this asymmetry to the farmers' disadvantage. Second, it highlights the particular type of agriculture that the use of AI tends to promote, i.e., big industrial monoculture, and the concerns that it raises.

[793] Rodrigues, Rowena and Jansen, Philip, "Brief report of the SIENNA foresight workshop on the social and ethical issues of AI and robotics", SIENNA, January 2019.
[794] Reuters, op. cit., April 2012.
[795] Wright and Schultz, op. cit., 2018.
[796] Trentesaux, and Rault, op. cit., 2017.
[797] Acemoglu and Restrepo, op. cit., Jan 2019, p. 4.
[798] Frey and Osborne, op. cit., Sept 2013, p. 13.
[799] Fleming, Aysha, et al., "Is Big Data for Big Farming or for Everyone? Perceptions in the Australian Grains Industry," *Agronomy for Sustainable Development*, Vol. 38, No. 24, 2018; Wolfert, Sjaak, et al., "Big Data in Smart Farming - A Review," *Agricultural Systems*, Vol. 153, 2017.

Firstly, deployment of AI in the agricultural sector risks further increasing the power imbalance between powerful agribusinesses and farmers.[800] This would exacerbate the latter's dependency on the former. In turn, this may endanger further the farmers' autonomy and freedom in their work. There are mainly two facets to this risk, one related to the machinery equipped with AI and the other related to data generated by AI systems. Regarding the former, (a) companies providing agricultural machinery (such as John Deere) have put in place contracts that forbid users to repair their equipment as, they claim, this would violate intellectual property rights.[801] Hence, farmers' "ownership and control over agricultural production" is being "expropriated from farmers and diverted to corporations".[802] This legal regime limiting farmers' control over the technology generates greater dependency on the technology providers and a loss of autonomy and agency on the part of the farmers as they are restricted from choosing and/or using their own repair agents.[803] This issue is further reinforced by the significant surveillance power of the AI technology providers over the farmers.[804] (b) The power asymmetry between the farmers and the AI providers (i.e., the big agribusinesses) is also affected deeply by the data economy at stake. The digitalisation of farming implies massive collection of data. While this data is generated by farmers on their land, companies processing it claim to own this data and require farmers to pay to gain access to it.[805] For instance, data collected by John Deere agricultural machinery are not openly accessible to farmers. This has been called the "'big data divide' between people and their data: they are rarely granted access to their own data, and they lack the tools or the context to analyse it – it is corporations, not individuals, that benefit from big data collection."[806] Hence, as Bronson and Knezevic put it, big data "has the potential to wade in on long-standing relationships between players in food and agriculture (e.g., between farmers and agricultural corporations)."[807]

The second set of ethical issues raised by AI in the agricultural sector resides in how it contributes to reinforcing a particular type of agriculture that has been shown to be problematic, i.e., big industrial monocultures. This is more an issue with the way that the technology is being implemented than with the technology itself. Indeed, Carbonnel notes that, although big data technologies could be "very useful for non-industrial farming practices, […] at present big data and data analytic tools are designed by big agribusinesses for industrial agriculture."[808] As Bronson and Knezevic observe, big data tools have the great potential to "normalise hegemonic farming systems".[809] Behind this model of farming, there is also a particular approach to farming that is being promoted, a highly rationalised and standardised one.[810] This is clearly illustrated by the quote "good farmers do not follow their gut, they follow data".[811] The "gut" in this quote actually corresponds to what is often a highly subtle and sophisticated knowledge developed over long periods of time by farmers on their soil, considering the local climate, and methods to ensure good production. These skills and knowledge may be disregarded

[800] It could be argued that the opposite would be true if AI systems were to be cheap and widely available; however, this is not the way AI in the agricultural sector is developing.
[801] Carbonell, Isabelle M., "The Ethics of Big Data in Big Agriculture," *Internet Policy Review*, Vol. 5, No. 1, 2016.
[802] Pechlaner quoted in Carbonell, op. cit., 2016, p. 5.
[803] Bronson, Kelly, and Knezevic, Irenam "Big Data in Food and Agriculture," *Big Data & Society*, 2018, p. 2.
[804] Carbonell, op. cit., 2016, p. 2 and 6.
[805] Bronson and Knezevic, op. cit., 2018, p. 1.
[806] Carbonell, op. cit., 2016, p. 3. The concept of 'big data divide' is by Mark Andrejevic.
[807] Bronson and Knezevic, op. cit., 2018, p. 1.
[808] Carbonell, op. cit., 2016, p. 4.
[809] Bronson and Knezevic, ibid., p. 3.
[810] Bronson and Knezevic, op. cit., 2018, p. 3.
[811] Carolan 2015 quoted in Mark Ryan, "Ethics of Using AI and Big Data in Agriculture: The Case of a Large Agriculture Multinational," *ORBIT Journal*, Vol. 2, No. 2, 2019, p. 6. https://doi.org/10.29297/orbit.v2i2.109

and eventually be lost in an agricultural sector increasingly captured by data-led farming that is controlled by big tech companies. This is socially and ethically problematic as it implies homogenisation and standardisation both of farming skills and products. Furthermore, industrial monoculture farming, i.e., the agricultural practice of massively producing a single crop or livestock species, has been shown to be quite problematic environmentally as it impoverishes soil and destroys ecosystems.[812] Hence, though the argument of the need to improve productivity because of the rising global population has validity, the capacity to do that through this type of farming is questionable considering how it can only last so long. In other words, it is a short-term productivism while current environmental challenges make it clear that we need to work toward sustainability.

7.2. Ethical issues with robotics applications

This subsection identifies and describes the ethical issues that may occur in various important application domains of robotics technology. It discusses, in turn, the issues in *transportation* (subsection 7.2.1), *law enforcement* (subsection 7.2.2), *defence* (subsection 7.2.3), *infrastructure* (subsection 7.1.4), *healthcare* (subsection 7.1.5), *companionship* (subsection 7.1.6), *manufacturing* (subsection 7.1.7), *exploration* (subsection 7.1.8), *service sector* (subsection 7.1.9), and the *environment* (subsection 7.1.10). Table 11 below lists the most important ethical issues that have been identified for each of these application domains.

Domain	**Ethical issues**	
Transportation	**- Safety** **- Security**	**- Transparency** **- Responsibility and accountability**
Law enforcement	**- Surveillance and privacy** **- Dehumanising of policing activity** **- Harms to communities** **- Robot control over humans** **- Responsibility and accountability**	**- Fairness (robots with weapons)** **- Safety (robots with weapons)** **- Security** **- Bias and discrimination**
Defence	**- Threat of uncontrolled escalation** **- Just war compliance**	**- Responsibility and accountability** **- Military virtue**
Infrastructure	**- Privacy**	**- Job losses**
Healthcare	**- Patient privacy and confidentiality** **- Quality of care** **- De-skilling of medical staff**	**- Patient integrity** **- Humanity in patient care**
Companionship	**- Security** **- Safety** **- Privacy** **- Effects on human sociality**	**- Effects on child care** **- Effects on elderly care** **- Issues with sex robots & romance**
Manufacturing	**- Safety** **- Job losses**	**- Quality of work** **- Responsibility and accountability**
Exploration	**- Environmental harms**	**- Interplanetary contamination**

[812] Union of Concerned Scientists, “Industrial Agriculture” https://www.ucsusa.org/our-work/food-agriculture/our-failing-food-system/industrial-agriculture

Service sector	**- Risks of robot autonomy** **- Job losses**	**- Robot control over humans**
Environment	**- Environmental harm** **- Animal wellbeing**	**- Responsibility and accountability** **- Privacy**

Table 11: Overview of ethical issues in major robotics application domains.

7.2.1. Transportation

One of the most common ethical problems encountered in discussions on autonomous vehicles (AVs), especially those sharing the roadways or airspace with human operators, is that of *forced choice decisions*. Otherwise known as *collision ethics* or *crash ethics*, forced choice decisions focus on how to program autonomous vehicles to "decide" between two or more unideal choices. For example, if there is a pedestrian in the way of a vehicle, and the only way to avoid hitting the pedestrian is to veer into a pylon, thus potentially killing the driver, which course of action should the car be programmed to take?[813] Debates on this topic mainly centre on what ethical approach to program autonomous vehicles with (e.g., deontology, consequentialism).[814,815,816] Should the car save the most people in any situation? Should the car prioritize the driver? Should the car be programmed to prioritize children? Or should it risk the driver's live to avoid hitting animals? These types of questions are at the heart of crash ethics.

Alongside applied ethical dilemmas such as collision ethics are those about normative questions of trust and accountability. If one designs a vehicle using an ethical approach that does not prioritise the safety of the vehicle's occupants, would anyone trust such a vehicle enough to purchase it? Furthermore, due to the decision-making process of AVs being rather opaque, it may be difficult to foster trust that the vehicle would make the very same decisions a human driver would generally make, even if the AV is better equipped to make decisions from a technical point of view (e.g., it has faster reaction times and increased awareness).[817] In addition, there are questions about responsibility and accountability. When an AV causes an accident, crashes, or harms another person, who is responsible and liable? (See also the part on "Responsibility and accountability" in subsection 5.1.3.) Discussions grappling with the issues of transitioning from human agents as responsible parties towards hybrid responsibility (human and AV) or to AV responsibility are focused on how to reform the legal and institutional aspects of transitioning to AVs, rather than designing the car to account for these changes.[818]

One of the other major, ongoing ethical discussions on AVs is that of privacy and data management. In order to make decisions, especially in cases of forced choices or involving pedestrians, AVs will need to have a plethora of sensors to assess the surrounding environment. Some of the unknowns

[813] For more such examples, see MIT's Moral Machine project: moralmachine.mit.edu/ (accessed 23 June 2018)

[814] Steinfeld, Aaron, "Ethics and Policy Implications for Inclusive Intelligent Transportation Systems," *Robotics Institute at Carnegie Mellon University*, 2010.

[815] Goodall, Noah J., "Ethical Decision Making During Automated Vehicle Crashes," *Transportation Research Record: Journal of the Transportation Research Board*, 2014.

[816] Bonnefon, Jean-François, Shariff, Azim & Rahwan, Iyad, "Autonomous Vehicles Need Experimental Ethics: Are We Ready for Utilitarian Cars?" 2015.

[817] Holstein, Tobias, Dodig-Crnkovic, Gordana & Pelliccione, Patrizio, "Ethical and Social Aspects of Self-Driving Cars," *Cornell University*, 2018.

[818] Fleetwood, Janet, "Public Health, Ethics, and Autonomous Vehicles," *American Public Health Association*, 2017.

surrounding this topic are what type of data the cars will collect, how long they will keep it, and to whom will the data be accessible. This is important as it is not only the "owner" of the vehicle who is generating the data for the vehicle; other passengers, pedestrians, and drivers are doing so as well. Thus, even if the owner agrees to some of these data monitoring practices, there are still many parties to be concerned about. Also, the distinct lack of transparency behind the decision-making algorithms of these vehicles, raises question about what data is generated exactly, and who has access to it. This brings up the other key point of security, as it is uncertain how "hackable" AVs will be and what type of damage security breaches could cause in these cases. The more institutions and commercial enterprises that have access to the data, the more potential entry-points into abusing these systems there are.[819,820,821]

When attempting to horizon scan for future ethical concerns of AVs, there seems to be a lot of focus on the adoption and widespread impacts of AV technology. For instance, once the use of AVs begins to grow to surpass that of regular vehicles, there will be many environmental, public safety, and traffic management/social order policies and practices that need to change to accommodate the change in use as well. For instance, will traffic flow be altered to accommodate different types of commuters as it currently is (carpool lanes, bus lanes, etc.)? Or will widespread use of AVs reduce pollution and help adopting nations reach sustainability and efficiency goals more easily?[822] Most importantly, how will regulators help transition users if the use of AVs begins to climb? This question is rather critical for future-planning in ethics for AVs, as once AVs become widely adopted, those who continue to drive manually become even larger risks to third parties as they become even more difficult to predict and do not have as many human drivers anticipating and accommodating their behaviours. Thus, there may reach a tipping point in which the ethical use of manual vehicles may no longer be justifiable in an environment that uses AVs as the majority or near majority.[823] As one may notice, these questions of the future are focused much more on an angle of social order surrounding AVs than on the AVs than on the AVs themselves. Highlighted in the ongoing problems of AVs, one could also speculate that future concerns will be those of increasingly complex decision-making algorithms, more attention to security breaches and data marketing, and more involved debates on the valuation of various human and non-human actors the car may come into contact with.

7.2.2. Law enforcement

In the law enforcement sector robots can be used for (1) surveillance and data gathering and processing, (2) handling potential explosives (reducing risks to human police personnel), and (3) handling weapons.[824] Proponents of the use of robots by law enforcement agencies (LEAs) argue they may improve the effectiveness and efficiency of policing activities and reduce costs by automating routine tasks. Robots are also of particular interest to LEAs as they can replace human officers in

[819] Fagnant, Daniel J. & Kockelman, Kara, "Preparing a Nation for Autonomous Vehicles: Opportunities, Barriers and Policy Recommendations," 2015.
[820] Holstein, Tobias, 2018, op. cit.
[821] Steinfeld, Aaron, 2010, op. cit.
[822] Mladenovic, Milos N. & McPherson, Tristram, "Engineering Social Justice into Traffic Control for Self-Driving Vehicles?" *Science and Engineering Ethics*, August 2016.
[823] Sparrow, Robert & Howard, Mark, "When Human Beings are Like Drunk Robots: Driverless Vehicles, Ethics, and the Future of Transport," *Transportation Research Part C: Emerging Technologies*, July 2017.
[824] Gettinger, Dan, and Arthur Holland Michel, *Law Enforcement Robots Datasheet*, Center for the Study of the Drone, Bard College, 2016; Asaro, Peter, "'Hands Up, Don't Shoot" HRI and the Automation of Police Use of Force", *Journal of Human-Robot Interaction*, Vol. 5, No. 3, 2016, pp. 55–56.

dangerous missions.[825] Though this is contested, some argue that robots may help de-escalate critical or tense situations as they may less be inclined to resort to force since they do not have the same need as humans do to defend themselves.[826]

However, the use of robots by LEAs raises a number of ethical issues. These are related to the particular "powers we entrust to the police", especially police's role of enforcing compliance to the law.[827] Asaro notes that, "[w]hile most of the recently developed police robots are remotely operated, rather than autonomous, and most are not weaponized, research continues into increasingly autonomous patrol robots with a clear potential for being weaponized."[828] It makes sense to distinguish between ethical issues related to the use of robots equipped with weapons and those related to the use of robots without weapons. Among the latter, key issues include the extended capacity for surveillance that the use of robots by LEAs enables, especially through the use of drones, and the threat to privacy this implies. The privacy issues this raises are similar to those identified in subsection 7.1.5 on the use of AI by LEAs, i.e., there is an expanded surveillance power that is both "wider and deeper".[829] Another ethical issue relates to the increasing removal of the human element from policing activities. This might lead to a form of policing that tolerates no exception, discussion, and negotiation between the police and the public and/or individual. In turn, this may further strain relations between the policing community and the public; this issue is even more critical for communities that already experience problematic relations with the police, e.g., certain ethnic minorities. As Joh observes: "Democratic policing involves trust and legitimacy, values that require human relationships. Robots should be a tool for safety, and not for further distancing."[830] Related to this issue, we may wonder, as Lin et al. do, whether there are "particular moral qualms with placing robots in positions of authority, such as police, prison or security guards, teachers, or any other government roles or offices in which humans would be expected to obey robots?"[831]

The second set of ethical issues relates to the use of weaponised robots by LEAs, i.e., robots equipped with lethal or non-lethal weapons. Experts and civil society organisations have expressed several concerns.[832] As Asaro argues, the design and use of robots to use force goes against Isaac Asimov's well-known first law of robotics according to which a "robot may not injure a human being or, through inaction, allow a human being to come to harm."[833] In the case of weaponised robots, the harm is caused intentionally. Although the difference between law enforcement and defence does imply

[825] For instance, a robot equipped with a bomb was used by Dallas police to kill a suspect in 2016. The justification for the use of the robot was that it would have been too dangerous for officers to go themselves to confront the suspect. See for instance: Graham, David A., "The Dallas Shooting and the Advent of Killer Police Robots", *The Atlantic*, 8 July 2016. https://www.theatlantic.com/news/archive/2016/07/dallas-police-robot/490478/

[826] Welinder, Yana, "Police Robots Could Reduce the Use of Deadly Force", *The New York Times*, 14 July 2016. https://www.nytimes.com/roomfordebate/2016/07/14/what-ethics-should-guide-the-use-of-robots-in-policing/police-robots-could-reduce-the-use-of-deadly-force

[827] Joh, Elizabeth E., "Policing Police Robots", *UCLA Law Review Discourse*, Vol. 64, 2016, p. 521.

[828] Asaro, op. cit., 2016, p. 56.

[829] Brayne, op. cit., 2017, p. 979.

[830] Joh, Elizabeth E., "Police Robots Need to Be Regulated to Avoid Potential Risks", *The New York Times*, 16 November 2016. https://www.nytimes.com/roomfordebate/2016/07/14/what-ethics-should-guide-the-use-of-robots-in-policing/police-robots-need-to-be-regulated-to-avoid-potential-risks

[831] Lin, Patrick, Keith Abney, and George Bekey, "Robot Ethics: Mapping the Issues for a Mechanized World", *Artificial Intelligence*, vol. 175, 2011, p. 947.

[832] Amnesty International, "Autonomous Weapons Systems: Five Key Human Rights Issues for Consideration", 2015. https://www.amnesty.org/en/documents/act30/1401/2015/en/

[833] Quoted in Asaro, op. cit., 2016, p. 68.

distinct ethical issues in these two sectors, the removal of the human element in the use of force remains the heart of the issue in both sectors.[834] As Asaro notes, in order to be lawful in the policing context, the use of force must satisfy the following requirements: "1) it must be necessary to prevent an imminent grave bodily harm or the death of a person; 2) it must be applied discriminately ; 3) it must be applied proportionately; and 4) the use of force must be accountable to the public."[835] According to Asaro, robots are unable to properly satisfy these requirements because they lack the capacity of judgement to assess whether the necessary conditions are met in order to legally resort to force. Furthermore, the risk of hacking poses a particular threat when robots are used in the law enforcement sector, even more so if these are equipped with weapons.[836]

Finally, in the future, if predictive policing programs (see subsection 7.1.5) were used not only to provide guidance to LEAs but to actually conduct operations, issues of discrimination raised by these programs would be further exacerbated as automatically implemented by robot police officers.[837]

7.2.3. Defence

As mentioned in the AI section on defence (subsection 7.1.4), recent technological developments in the fields of AI and robotics in this sector have led to intense ethical, legal, and policy debates. This is due to the increased autonomy rendered possible by AI. However, because a number of these developments have been implemented with a physical component, this report examines these in the present robotics section. Although we recognise that some applications are software only, as mentioned in the AI defence section, it was decided to expand on the ethical issues they raise in the robotics section as the ethical debate explores AI and robotics issues together. It is necessary to introduce a number of distinctions within this debate; hence, in order to avoid repetitions, it was decided to do so in the robotics defence section only. Most of the ethical debates focus on AI and robotics technologies in relation to weapons. Key distinctions need to be made here in relation to where the autonomy intervenes as this has ethical implications.

Is the autonomy at the level of the critical functions of the decision to kill or not? As Asaro notes, "what makes a weapon autonomous is that the determination to use violent or lethal force has been made by an automated process, i.e., an algorithm."[838] If the autonomy does not intervene at this level, ethical issues at stake are not as critical. Righetti et al., note that "autonomy is becoming pervasive in noncritical components of weapon systems, such as transport, navigation, or surveillance, and has already had an impact on the use of military force by nations. Partial autonomy in the navigation and surveillance capabilities of drones, e.g., has been decisive in the rapid and extensive deployment of the controversial U.S. drone program".[839] According to Asaro, a number of AI applications in defence can be seen as being "reasonable", including "pattern recognition systems for filtering surveillance data, to software for blast damage assessment, to guidance systems on missiles, to mines and

[834] Graham, op. cit. 8 July 2016.
[835] Asaro, op. cit., 2016, p. 60.
[836] Joh, op. cit., Nov 2016.
[837] Joh, op. cit., 2016, p. 540.
[838] Asaro, Peter, "Algorithms of Violence: Critical Social Perspectives on Autonomous Weapons", *Social Research*, Vol. 86, No. 2, 2019, p. 539.
[839] Righetti, L., Q.-C. Pham, R. Madhavan, and R. Chatila, "Lethal Autonomous Weapon Systems", *IEEE Robotics & Automation Magazine*, March 2018, p. 124.

munitions that self-destruct or deactivate after a period of time."[840] It is nonetheless essential that, as it is the case in most use of AI, users of the technology, in this case the military personnel, be well trained on this technology to then be able to adequately understand the results obtained by the machines.

However, the discussion is radically different when it comes to autonomous weapons, i.e., when it is an algorithm that determines the use of violent or lethal force. Such use of AI technology is widely debated at the ethical, legal and policy levels. While some states hold the view that it is too early to start regulating these weapons, often called Lethal Autonomous Weapons (LAWs), other nations are calling for a ban.[841] The ethics of LAWs is also widely debated at the academic levels. Some, although very few, such as Ronald Arkin, argue that there is a moral duty to use LAWs in wars as they are, they claim, more ethical than humans.[842] For instance, they would not kill out of anger as humans would.[843] However, numerous experts and organisations[844] are strongly arguing against the development and use of LAWs on ethical and legal grounds. A key element in this debate is that these weapons would not be able to comply with the rules of wars – rules that have their roots in the ethics of war and just war theory – and are enshrined in International Humanitarian Law (IHL). Experts are especially concerned that LAWs are not able to comply with the key principles of distinction (distinction between legitimate targets and illegitimate targets, especially civilians) and of proportionality (avoidance of excessive force in relation to the military objectives). Opponents of LAWs argue that they would not be able to uphold these principles because their proper implementation requires the capacity to make moral judgement on the basis of an assessment of context, a judgment that machines cannot make according to them.[845]

Furthermore, there is the issue of the gap of responsibility and accountability that is brought to critical levels with LAWs considering that the concern life-and-death decisions. Another argument of the opponents of LAWs consist in challenging a key argument put forward by proponents of these weapons. According to the latter, LAWs would reduce the number of civilian casualties because of a more precise targeting. However, this utilitarian argument may be questioned for being short-sighted. An exploration of the potential deeper implications of LAWs for the conduct of war leads to more complex and concerning consequences.[846] A key element here is the increased speed of weapons that may lead to much more rapid conflict escalations that may have dramatic consequences for civilians. More generally, the deployment of LAWs could profoundly disrupt international relations in such a

[840] Asaro, Peter, "Why the world needs to regulate autonomous weapons, and soon", *Bulletin of the Atomic Scientists*, 27 April 2018. https://thebulletin.org/2018/04/why-the-world-needs-to-regulate-autonomous-weapons-and-soon/

[841] For a review of the positions of seven key countries in this debate, please see: Gronlun, Kirsten, "State of AI: Artificial Intelligence, the Military and Increasingly Autonomous Weapons", Future of Life Institute Website, 9 May 2019. https://futureoflife.org/2019/05/09/state-of-ai/

[842] Arkin, Ronald, *Governing Lethal Behavior in Autonomous Robots*. Boca Raton, Chapman and Hall/CRC Press, 2009.

[843] Righetti, L., et al., op. cit., 2018, p. 125.

[844] These organisations include in particular Human Rights Watch, the International Campaign to Ban Killer Robots, and the International Committee of the Red Cross.

[845] Asaro, Peter, "On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making", *International Review of the Red Cross*, Vo. 94, 2012, pp 687-709. Roff, Heather and Richard Moyes, "Meaningful Human Control, Artificial Intelligence and Autonomous Weapons" Briefing paper prepared for the Informal Meeting of Experts on Lethal Autonomous Weapons Systems, UN Convention on Certain Conventional Weapons, April 2016.

[846] Asaro, "Algorithms of Violence: Critical Social Perspectives on Autonomous Weapons", op. cit. 2019.

way that they have grave effects on international security. Finally, many experts and lay persons are concerned about the morality of delegating the decision to kill to a machine.

7.2.4. Infrastructure

The most popular robots currently in use for infrastructural applications are unmanned aerial vehicles (UAVs), which are used mostly for inspection work. Unfortunately, the ethics literature on use of such robots in infrastructure is rather sparse. One reason for this could be that the implementation and application of such robots has taken longer, presented more problems, and become more expensive than initial projections had anticipated,[847] thus stalling subsequent research on the ethical problems that are currently going on and may happen in the future. Based on the current and future applications discussed in subsection 3.2 of the SIENNA D4.1 report on this topic,[848] however, a few ethical problems can be speculated at by examining civil applications of drones and smart city debates.

One of the biggest concerns here is the observation of spaces frequented by human beings and its effects on privacy. As infrastructural robots are doing their jobs, what data is being captured, how is it used, and who is in charge of maintaining, implementing, and safeguarding the data collected?[849,850] Even in infrastructure applications of robots that have nothing to do with surveillance or interacting with humans, personal data is still being collected and still raises questions about identifiability, data protection, and equality.[851] Since it is envisioned that many UAVs will be used in the upkeep, monitoring, and implementation of systems in smart cities, many concerns relating to UAVs are also reflected in smart city discussions—data access and control, the erosion of privacy in public spaces, et cetera.[852]

A further set of ethical concerns with infrastructural applications of robots relate to job security and safety. How many human workers will be losing their jobs as a result of the use of such applications of robots? Should the focus of infrastructure robots be on collaboration rather than replacement? Or should it only be on jobs that are deemed too risky for human workers?

7.2.5. Healthcare

Currently, there are two main groups of ethical issues with the use of robots for healthcare: concerns about privacy and concerns about responsibility. Discussions on privacy relate to a need for care robots and surgical robots to adhere to already-present legal and ethical frameworks present for human carers on this topic. Furthermore, since robots are capable of acquiring, storing and sharing a larger quantity and variety of data than human carers, more attention needs to be placed on how to protect

[847] Edlich, Alex & Sohoni, Vik, “Burned by the Bots: Why Robotic Autonomation is Stumbling”, *McKinsey & Company*, May 2017.
[848] Jansen, et al., 2018, op. cit.
[849] Jensen, Ole B., “Drone City—Power, Design, and Aerial Mobility in the Age of Smart Cities”, April 2016.
[850] Finn, Rachel & Wright, David, “Privacy, Data Protection and Ethics for Civil Drone Practice: A Survey of Industry, Regulators and Civil Society Organisations”, *Computer Law & Security Review 32*, 2016.
[851] Finn, Rachel & Donovan, Anna, “Big Data, Drone Data: Privacy and Ethical Impacts of the Intersection Between Big Data and Civil Drone Deployments”, *The Future of Drone Use*, October 2016.
[852] Kitchin, Rob, “The Real Time City? Big Data and Smart Urbanism”, *GeoJournal 79*, November 2013.

patients from data hacking, exploitation, and data acquisitions they did not consent to. Privacy concerns also encompass transparency concerns about data ownership and viewership.[853,854,855,856]

In relation to responsibility, questions are raised as to who is responsible if a care robot unintentionally harms a human? Furthermore, who is responsible for the maintenance of the care robot and the answering of technical questions from patients that use the robot? Will the use of robots be subject to high insurance premiums like the services provided by physicians and other healthcare practitioners? Especially in cases of healthcare robots being employed outside of the hospital, these questions are of paramount concern to ensure appropriate preventative frameworks and follow-up if and when problems are faced with healthcare robots.[857,858,859]

Other, less-explored ethical issues with healthcare robotics include potential negative effects on quality of care and patient integrity,[853] and concerns about machine reliability,[860,861] and a potential de-skilling of medical staff.[862] Quality of care and patient integrity are about how to best design robots that will treat patients with compassion, dignity and respect.[863] Machine reliability is needed to ensure safety and cultivate trust: What safeguards need to be put in place, especially with collaborative robots used in surgery or for caregiver cooperative actions, to ensure that these robots perform reliably enough to trust the delegation of certain highly-sensitive operations and vulnerable groups?[864,865] De-skilling among medical staff may result from, for example, surgical robots that prevent surgeons from keeping their skills up to date through daily practice. De-skilling can become a problem for surgeons when a complex or emergency procedure requires manual intervention.

Quality of care and reliability may be important issues with the increasing shift towards the “hospital at home”, where care robots and other technologies, increasingly find their way into patients’ own homes.

As the use of healthcare robots becomes more widespread and their capabilities increase, much of the debate on possible future ethical issues surround impacts on human-to-human relations and job replacement. Many researchers speculating in this area are concerned that the lack of human contact with the use of robotic caregivers, particularly in homecare, may lead to vulnerable individuals (elderly people, children) becoming more isolated and feeling detached from their communities. Thus, many of these questions surround how to supplement the convenience and cost-reduction of care robots with the human touch of a person. Should care robots be supplemental assistants? Should only certain, less vulnerable persons be able to use care robots full-time? Are there ways to design care robots so

[853] Van Wynsberghe, Aimee, *Healthcare Robots*, 2015.
[854] Stahl, Bernd Carsten & Coeckelbergh, Mark, “Ethics of Healthcare Robotics: Towards Responsible Research and Innovation”, *Robotics and Autonomous Systems, 86*, December 2016.
[855] Lutz, Christoph & Tamò, Aurelia, “Privacy and Healthcare Robots—An ANT Analysis”, 2016.
[856] Mavroforou, A., Michalodimitrakis, E., Hatzitheo-Filou, C., & Giannoukas, A. “Legal and Ethical Issues in Robotic Surgery”, *PubMed*, February 2010.
[857] Van Wynsberghe, Aimee, 2015, op cit.
[858] Ibid.
[859] Easton, Catherine, “Carry on Automat(r)on: Legal and Ethical Issues Relating to Healthcare Robots”, *Tech Law*, May 2013.
[860] Stahl, Bernd, 2016, op. cit.
[861] Mavroforou, A., 2010, op. cit.
[862] The included of de-skilling as an issue is based on comments on a draft version of this report.
[863] Van Wynsberghe, Aimee, 2015, op cit.
[864] Ibid.
[865] Simpson, Trudy, “Rise of the Healthcare Robots: Five Ethical Issues to Consider”, *Christian Medical Fellowship*, March 2016.

as to reduce the impacts of the loss of human contact?[866,867,868] Bridging onto this is the ubiquity of decisions made by the robot and the information available to it. Can and should care robots override patients' desires?[869,870] Should robots be able to deceive patients if it is for their own good ("tricking" them into taking necessary medicines or exercising)? To what extent should the robot be responsible for the patient? Is it required to be a companion and a caregiver?[871,872] Should robots filter out information not pertinent to the patient's healthcare?[873] How do we ensure present inequalities in care and treatment are not perpetuated with care robots?[874]

The other main point of concern for ethicists and researchers in this field is that of job replacement. On one side of the debate, briefly outlined above, is the question of whether robots can sufficiently replace human caregivers in certain contexts, or would be desirable replacements in particular contexts.[875,876,877] This is more due with their capabilities and performance in these particular roles. On the other side of the debate is the question of what to do with the various human beings that will be out of work if the use of these robots becomes more desirable than they are for various reasons not limited to cost, efficiency, quality of care, and effectiveness.[878,879] Would it perhaps be preferred to refocus on collaboration rather than on replacement? Or should we focus on using robotics to enhance and supplement human beings in a more rehabilitative stance? And if robots are going to replace humans in caregiving, to what extent is it in favour of the patient to know that they are dealing with a robot? Humanoid robots may (in the future) be able to deceive children or elderly people by disguising themselves as human caregivers in order to provide better care or to build a better relationship with patients who are sceptical about dealing with robots. What consequences will this have on the relationship between patient and caregiver and responsible family member, for example, on trust?

7.2.6. Companionship

Current ethical concerns about the application of companion robots are not so different from those of any other more ubiquitous technology discussed above: security, privacy, and safety again are central topics. If these robots are going to be in constant or near constant contact with their human companions, they are privy to a level of intimacy and information inaccessible to other technologies—consider sex robots, robot nannies, or companion robots for the elderly people. All of these examples are used either in vulnerable relational contexts or with vulnerable user groups that stand much to

[866] Ibid.
[867] Van Wynsberghe, 2015, op cit.
[868] Stahl, 2016, op. cit.
[869] Van Wynsberghe, 2015, op cit.
[870] Stahl, 2016, op. cit.
[871] Ibid.
[872] Simpson, 2016, op. cit.
[873] Lutz, 2016, op. cit.
[874] Van Wynsberghe, 2015, op cit.
[875] Ibid.
[876] Stahl, 2016, op. cit.
[877] Simpson, 2016, op. cit.
[878] Van Wynsberghe, 2015, op cit.
[879] Stahl, 2016, op. cit.

lose from security breaches, hacking of the robots, or privacy violations. As such, it is no surprise that these ethical concerns are of top priority for roboticists and designers alike.[880,881,882,883]

Other important ethical issues concern the impact on human relations: How will companion robots change how humans interact with other humans? Authors working from this angle are concerned that companion robots may lead to social isolation and increased objectification of human beings as individuals may grow to prefer the "easier" relation of the companion robot. In other cases, it may be that companion robots are able to provide a sense of social interaction for user groups that are already becoming socially isolated (older adults, individuals with social anxiety) to help overcome this problem.[884,885,886] Another concern is that of deception: Is there something problematic about making machines that encourage users to empathize, relate-to, trust, and emotionally invest in them without the companion bots actually being capable of reciprocating these values and expectations? Does such a unidirectional relationship with robotic systems stunt human-to-human relations in any way?[887,888,889]

Another, more popular ethical concern when it comes to companion robots is of the types and design of sex robots that should or should not be created. Should child sex robots or rape robots be commercialized and available for widespread use?[890] Should robots be used in brothels or in other areas of prostitution? Further types of discussions surround topics about how companion robots should be designed: Should humanoid robots be designed in general? If so, should robots be designed with human genders and be anatomically correct?[891]

Although most of the dialogue surrounding companion robots is focused on addressing and keeping up with current uses and concerns, there are a few possible directions that these conversations can turn towards in the future. To begin, security, privacy, and safety concern may not go away as companion robots become more advanced, but rather become more prevalent as companion robots function according to more advanced hardware and software. Other likely types of questions may concern the appropriate contexts and uses of companion robots. Are there types of robots that cannot be companions or pets? Is it ethical to make a humanoid robot that is intended for use as a pet? Or to make a pet programmed with a human-like AI? Further, are there areas and aspects of life where companion robots should not be used (e.g., for child care, for people with cognitive impairments, for lifeguards)? Robot slavery may also be a point of concern in the future, but more about this will be discussed in the service sector subsection.

[880] Bisconti Lucidi, Piercosma & Nardi, Daniele, "Companion Robots: The Hallucinatory Danger of Human-Robot Interactions", *Association for the Advancement of Artificial Intelligence*, 2018.

[881] UNESCO, "Section 3.4.4 Companion Robots", *Report of Comest on Robotics Ethics*, September 2017.

[882] Sharkey, Noel & Sharkey, Amanda, "The Crying Shame of Robot Nannies: An Ethical Appraisal", *Interaction Studies*, 11(2), 2010.

[883] Veruggio, Gianmarco & Operto, Fiorella, "Roboethics: Social and Ethical Implications of Robotics", *Springer Handbook of Robotics*, 2008.

[884] Bisconti, 2018, op. cit.

[885] Simon, Matt, "Companion Robots are Here. Just Don't Fall in Love with Them", *Wired*, February 2017.

[886] Bali, Meghna, "Companion Robots: What are the Ethical Implications of Intimate Human-Machine Relationships?", *ABC News*, August 2017.

[887] Bisconti, 2018, op. cit.

[888] Simon, 2017, op. cit.

[889] Dumouchel, Paul & Damiano, Luisa, *Living with Robots*, 2017.

[890] Foundation for Responsible Robotics, "Our Sexual Future with Robots", 2017.

[891] Pearson, Yvette & Borenstein, Jason, "Creating 'Companions' for Children: The Ethics of Designing Esthetic Features for Robots", *AI & Society*, February 2014.

7.2.7. Manufacturing

Some of the most important present concerns of robots in manufacturing are safety risks to people. These risks can originate in a malfunction due to engineering or human errors. Employees are often working closely and intensively with the industrial robots, making them vulnerable to overestimating their abilities. One of the riskiest situations involves errors in human judgment. In this case, personnel are getting too comfortable around the robots and trust themselves to know its predictable motions and therefore place themselves in dangerous positions while supervising, operating or maintaining the robot.[892]

A few important potential future concerns for industrial robots originate from developments that will enhance the flexibility and context awareness of these robots. The role of the industrial robot is changing. While robots are currently still controlled by human operators and function as assistive tools, in the near future industrial robots may take the role of a collaborative co-worker. This shift in role changes the human-robot relationship which will have consequences in terms of responsibility for the robot's actions, safety regulations and design strategies for the industrial robots.[893]

Increasing transparency and defining responsibilities when it comes to the use and maintenance of the robots are of ethical concern for the future. At present, there seems to be a lot of uncertainty about who is responsible for the robot's actions. For example, if a robot harms a person, is it the manufacturer or the company that employs the robot? Regarding the robot's behaviour, there are calls for increased transparency so that human users are kept "in-the-loop" about the robot's decisions and action patterns. Importantly, it might be a problem if the robot's operation is so complex and highly technical that workers are unable to understand the robots' actions and accommodate it through their own practices and behaviours.

Finally, the last ethical concern to be discussed is that of training advancements and floor management both with the current and future role of the industrial robot. What types of training needs to be completed to ensure workers are psychologically and physically prepared for collaborating with robots in industrial contexts? Are there different types of hiring practices that will need to be used? What about organization of new human-machine assemblages?[894,895] The change of robots being assigned as a co-worker facilitates the possibility of using robots to replace the human labour within the manufacturing industry, while the implementation of robots does not necessary lead into a net difference in jobs, the low skilled jobs however will be the main victim, affecting only certain, already vulnerable demographic groups.[896]

7.2.8. Exploration

The ethical impacts of robots that are created purely for exploration are not yet widely considered in the literature. In many cases, exploration seems to relatively low impact as long as the missions are kept purely to discovery and data collection in locations on Earth (e.g., the deep sea). One thing to

[892] Murashov, Vladamir, Hearl, Frank & Howard, John, "Working Safely With Robot Workers: Recommendations for the New Workplace", *Journal of Occupational and Environmental Hygiene*, *13*(3), 2016.
[893] Ibid.
[894] Fletcher, S.R. & Webb, P., "Industrial Robot Ethics: Facing the Challenges of Human-Robot Collaboration in Future Manufacturing Systems", *A World with Robots: International Conference on Robot Ethics 2015*, 2017.
[895] Francis, Sam, "Robot Ethics: Three Things Industry Can Learn from New Robotic Standards", *Robotics & Automation News*, March 2017.
[896] West, Darrell M., "What Happens if Robots Take the Jobs? The Impact of Emerging Technologies on Employment and Public Policy", *Centre for Technology Innovation at Brookings,* October 2015.

always keep in mind for these types of robots is the potential for disturbing the locations and the networks of inhabitants, as well as being considerate of the amount and disposal of electronic waste generated when missions are unsuccessful or the robots no longer have a use.

As for exploration outside Earth, on other celestial bodies (i.e., on extra-terrestrial planets or moons), there is the potential for (biological) *interplanetary contamination* by space probes or spacecraft. There are two kinds of interplanetary contamination: *forward contamination* and *backward contamination*. Forward contamination, which involves the transfer of life and other contaminants from Earth to another celestial body, may carry extra-terrestrial planetary protection concerns. Backward contamination, which involves the introduction of extra-terrestrial organisms and contaminants into Earth's biosphere, may carry safety concerns for human beings and the environment on Earth.

An example of measures to prevent forward contamination is the US space agency NASA's effort to destroy its Cassini probe at the end of its mission by directing it to enter Saturn's atmosphere, thus preventing the possibility of it contaminating Saturn's moons.

7.2.9. Service sector

Service sector and companionship applications of robots possess many overlaps. In fact, nearly all questions and concerns raised in one area could be asked of the other. This seems to be the case in fields where robots are taking on performative roles in place of humans, rather than enhancing or extending human capabilities or capacities, like in industrial or healthcare applications. Furthermore, since these types of robots are more involved in more relational and intimate areas of human life, such as caregiving, physical intimacy, or aid, human beings depend and relate to these robots rather differently than they do to robots that are seen to be of more apparent instrumental value.

One of these dominating ethical concerns that overlap, particularly highlighted in the service industry, is the question of robot autonomy. To what extent should robots be programmed to make decisions without human approval or interference? What are acceptable value trade-offs in the pursuit of more automated service? For example, is it desirable to sacrifice privacy for convenience? Security for ubiquity? Transparency for efficiency? To program these robots to serve humans remotely effectively, they need to be programmed with data containing preferences, and potentially containing biases and stereotypes, in order for machine learning to take place and the machines to be adaptable enough to be useful rather than more of a hassle than human service members.[897] As such, many reoccurring arguments as mentioned in the Security segment of this section can be revisited again here—control, accountability, and transparency being top concerns.[898,899]

As for future concerns with service sector robots, being aware of causing human job losses and the impact on the service industry is paramount for ensure these robots have a positive impact on the service industry as well. This may be good reason to consider service robots in a role more akin to co-bots than as a replacement for human service workers. A 2018 report on human rights cautions the widespread replacement of human workers as conducive of exploitative environments and fear that human workers will have to enter into dangerous, undesirable, or potentially abusive work environments in order to keep a job at all.[900] Thus, part of being morally aware of robots in this way

[897] Riek, Laurel & Howard, Don, "A Code of Ethics for the Human-Robot Interaction Profession", *We Robot*, 2014.
[898] Tamburrini, Guglielmo. "Robot Ethics: A View from the Philosophy of Science", *Ethics and robotics*, 2009.
[899] Veruggio, et al., op. cit.
[900] Verisk Maplecroft, *Human Rights Outlook 2018*, July 2018.

also comes from being mindful of their use in relation to humans, and focusing on coexistence and collaborative efforts between humans and robots in the areas of service and companionship particularly.[901] Furthermore, although it is not commonly mentioned, there have been a few authors who argued that using humanoid robots in certain service contexts bears striking parallels to slavery, and may prove to become even more problematic as these robots and their capabilities become more advanced.[902,903]

7.2.10. Environment

Environmental robots can be split up in three different domains: (1) robots *in* ecology, used for environmental research applications (such as drones and UAVs to monitor the environment and count species[904]); (2) robots *for* ecology, used to specifically carry out environmental research, which are thus a subsection of the former (e.g., bio-mimicking robots for bacterial locomotion and robots that climbs and inspects trees); and (3) robots that enforce or control environmental or ecological factors (e.g., robots to clean contaminated water).

While the notion of using environmental robots usually generates positive support, there may be unexpected drawbacks that cause ethicists to weigh the risks and benefits more carefully. For instance, using undisguised drones to monitor an endangered species could increase stress levels of the animals. With the right design, however, robots have the potential to be far less invasive than the presence of a human field researcher.[905] One more difficult to assuage concern would be that the crashing down of a drone could cause environmental degradation through toxic and unrecoverable debris.

Moreover, following a similar discussion in subsection 7.2.4, even when UAVs are used to monitor the environment, they may still be collecting data on human beings. As such, issues of privacy, data protection and transparency are still relevant even here.[906] Furthermore, scientists who use environmental robots also have responsibilities over their secondary effects. For instance, if in the process of investigating wildlife, they stumble upon new data signalling threats to an ecosystem, then scientists should be responsible for sharing that data as well. Ethical concerns about robot dependence may also arise if ecosystems and environments become reliant upon robotic assistance. If robots are used to fill ecological gaps, who is responsible for their maintenance and continued contribution if the environments cannot exist without them? This stress on maintenance responsibility gives further cause to consider the robots' designs, materials, and product lifespan. While there are concerns about the use of inorganic materials in the natural environment, there may be even more significant concerns on how organically engineered material will affect its surrounding ecosystem in both the short and long term.[907]

[901] van Wynsberghe, Aimee, "Service Robots, Care Ethics, and Design", *Ethics and Information Technology 18*(4), December 2016.

[902] Singler, Beth, "Are We Expecting Automation to Give Us Modern Day Slaves?", *World Economic Forum*, May 2018.

[903] Brooks, Victoria, "Samantha's Suffering: Why Sex Machines Should Have Rights Too", *The Conversation*, April 2018.

[904] Ivošević, Bojana, Han, Yong-Gu, Cho, Youngho & Kwon, Ohseok, "The Use of Conservation Drones in Ecology and Wildlife Research", *Ecology and Environment*, *38*(1), February 2015.

[905] Ivošević, Bojana, Han, Yong-Gu, Cho, Youngho & Kwon, Ohseok, "Monitoring Butterflies With an Unmanned Aerial Vehicle: Current Possibilities and Future Potentials", *Journal of Ecology and Environment 41*(1), 2017.

[906] Finn, Rachel, and Anna Donovan, 2016, op. cit.

[907] Van Wynsberghe, Aimee & Donhauser, Justin, "The Dawning of the Ethics of Environmental Robots", *Science and Engineering Ethics 24*(6), December 2018.

7.3. Ethical issues for different types of users and stakeholders

In this section, we review and discuss ethical issues that affect different stakeholder categories. We consider how both end users and other stakeholders of AI and robotics are affected by the introduction and use of these technologies, and what ethical issues are raised. We consider the following demographic categories: *gender* (subsection 7.3.1), *race and ethnicity* (subsection 7.3.2), *age* (with a focus on children in subsection 7.3.3 and a focus on the elderly in subsection 7.3.4), *ability* (with a focus on people with mental and physical disabilities in subsection 7.3.5), and *educational level* and *income level* (both in subsection 7.3.6).

7.3.1. Gender

In relation to gender and AI and robotics, ethical issues have been raised with respect to employment, bias in design, and the lack of women in the technology sector. Starting with employment, we refer back to the discussion of gender in our discussion of mass unemployment and AI in section 5.2.2. There, we claimed that studies do not agree on the impact of automation along gender lines. We cited a study of AI automation and US employment by Muro, Maxim and Whiton (2019),[908] who find that men are more at risk to lose their job due to automation than women, 43% to 40%, due to their overrepresentation in manufacturing, transportation and construction jobs that are at risk for automation, and due to the overrepresentation of women in occupations in sectors like health care, personal services, and education that are relatively safe. In contrast, the World Economic Forum (2018) has found that 57 percent of jobs at risk for disruption belong to women.[909] They take into account that, according to their analysis, at-risk jobs in professions dominated by men have more reskilling and job transition options than those in professions dominated by women. Other studies of the impact on employment of AI and robotics automation also show mixed results, so it is as yet unclear whether men or women will be more affected.

Turning now to gender bias in design, there is much more agreement between studies: AI and robotics technologies often contain gender biases and display stereotypes, and they do so to the disadvantage of women. We will review three specific issues: algorithmic gender bias, genderedness in the usability of AI and robots, and gender stereotyping in robots and intelligent virtual assistants. Algorithmic gender bias, to begin with, is a specific type of algorithmic bias, as discussed in the subsection on justice and fairness in section 5.1.3 of this report. It is bias in the treatment of individuals and social groups represented by the system or otherwise affected by the system's decisions or recommendations. An example of algorithmic gender bias is an AI system used by Amazon.com Inc. to review job applicant's resumes. A review of the system revealed that it systematically downgraded female applicants for technical posts because it drew from past hiring practices to predict success, and most past jobs had gone to men.[910]

[908] Muro, Mark, Robert Maxim and Jacob Whiton, "Automation and Artificial Intelligence: How machines are affecting people and places," *Brookings Institution*, 2019. https://www.brookings.edu/wp-content/uploads/2019/01/2019.01_BrookingsMetro_Automation-AI_Report_Muro-Maxim-Whiton-FINAL-version.pdf

[909] World Economic Forum, *The Global Gender Gap Report 2018.* Retrieved at http://www3.weforum.org/docs/WEF_GGGR_2018.pdf.

[910] Dastin, Jeffrey, "Amazon scraps secret AI recruiting tool that showed bias against women," *Reuters*, 10 October 2018. Retrieves at https://www.reuters.com/article/us-amazon-com-jobs-automation-in...-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

Genderedness in the usability of AI and robots is a related issue and concerns what was called functional bias in the subsection on justice and fairness in section 5.1.3. Functional bias implies that AI systems offer functionality that serves the interests of certain social groups of users more than those of other groups. Functional gender bias is therefore a form of bias in which gendered interests, goals, concerns, traits, abilities, roles or cognitive and behavioural styles of end-users are unequally supported by the system. A companion robot that either explicitly or implicitly assumes that humans they interact with are male, for example, displays functional bias. Another example of functional gender bias is found in the AI-powered targeted advertising system of an advertiser, which was shown to show job advertisements in science, technology, engineering and mathematics less frequently to women then to men.[911] Moving beyond algorithmic bias and functional bias, Adam (1998) has argued for the existence of more fundamental gender biases in AI that are based in the Cartesian, disembodied and decontextualized conception of rationality that is found in AI systems.[912]

Gender stereotyping in robots and intelligent virtual assistants is a phenomenon that has, like algorithmic gender bias, received much recent coverage, both in academic and popular media. Robots are often genderless, but when they have a humanoid appearance, they are often assigned a gender. Genderedness is indicated through appearance, voice, gendered answers and responses, and the robot's name. Intelligent virtual assistant, such as Siri, Cortona and Alexa, are gendered by voice and by name, as well as genderedness in some of their responses (particularly about themselves). Robertson (2017) has shown how Japanese male and female robots in different social roles display the same gendered patterns in the division of labour as do humans.[913] Alesich and Rigby (2017) point out that most virtual assistants developed in the U.S. have female voices and names, and claim that this suggests to users that personal assistants are women.[914] A search of images of "female robot" and "woman robots" (in August 2019) shows that the vast majority of both real and fictional female robots have shapely bodies with slender waists and large breasts. Robots and intelligent virtual assistants may in this way end up perpetuating gender stereotypes.

These gender biases and gender stereotypes in AI systems and robots are not unrelated to the final topic discussed here, namely, the lack of women in the AI and robotics technology sector. Studies have shown that only 22 percent of employees in AI are women.[915] In a recent international prize competition, the 2015 DARPA Robotics, 95 percent of the 444 participants were men.[916] In the European Union, only 16.3 percent of computer science students and only 17.2 percent of employed ICT specialists are women.[917] In the United States, only 18 percent of computer science bachelor graduates are women.[918] There has been much recent reporting, as well, on sexism and discrimination

[911] Lambrecht, Anja and Catherine Tucker, "Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads," *SSRN*, 2018, retrievable at https://ssrn.com/abstract=2852260 or http://dx.doi.org/10.2139/ssrn.2852260.

[912] Adam, Alison, *Artificial Knowing: Gender and the Thinking Machine*, Florence, KY, USA: Routledge, 1998.

[913] Robertson, Jennifer, *Robo Sapiens Japanicus: Robots, Gender, Family, and The Japanese Nation,* University of California Press, 2017.

[914] Alesich, Simone, and Michael Rigby, "Gendered Robots: Implications for Our Humanoid Future," *IEEE Technology and Society Magazine*, Vol. 36, No. 2, 2017, pp. 50-59.

[915] World Economic Forum, 2018, op. cit.

[916] McFarland, Matt, "DARPA's Robotics Challenge has a gender problem." *Washington Post*, June 5, 2015. Retrieved at http://www.washingtonpost.com/blogs/innovations/wp/2015/06/05/darpas-robotics-challenge-has-a-gender-problem/.

[917] Eurostat, "Girls and women under-represented in ICT," 25 April 2018. Retrieved at https://ec.europa.eu/eurostat/web/products-eurostat-news/-/EDN-20180425-1.

[918] National Science Board, *Science & Engineering Indicators 2018*. Retrieved at https://nsf.gov/statistics/2018/nsb20181/report.

in the technology sector. A 2015 survey of 200 senior-level women in Silicon Valley showed that 84 percent had been told they were "too aggressive" in the office, 66 percent reported being excluded from important events because of their gender, and 60 percent reported unwanted sexual advances in the workplace.[919] Todd (2015) also presents the argument that AI draws less women because the field currently de-emphasizes humanistic and communal goals.[920]

Various scholars have linked gender biases and stereotypes in AI and robotics to the underrepresentation of women in these fields.[921] Oudshoorn, Rommes and Stienstra (2004) have argued that the risk of having a homogeneous designer community (with little stakeholder engagement) is that designers tend to use the "I-methodology": a design practice in which designers consider themselves as representative of the users.[922] If designers are mainly men, it follows that the technology that is developed mainly reflects the needs, preferences, and attitudes of men. This presents a strong argument for diversification of the AI and robotics workforce, next to the adoption of user-centred design methods.

7.3.2. Race and ethnicity

In relation to race and ethnicity, several issues have been raised regarding AI and robots and unemployment, workforce and bias. A major, general concern regarding AI-induced decisions is that these systems operate as "black boxes" such that individual users are unable to understand why and how decisions have been made.[923] This non-transparency potentially leads to concerns about perceived racial discrimination on several different levels that will be discussed in what follows.

With regards to hiring decisions based on AI algorithms, a concern about racial discrimination comes up that is similar to the aforementioned bias in design concerning gender. When AI decisions are based on biased training data, for example if the training data is biased against a particular race, so will the decision based on that data be biased.[924] This might lead to disadvantages in the hiring process, and consequently to a higher unemployment rate. As pointed out in section 5.1.3 of this report, certain ethnic groups in the United States are more at risk of suffering from unemployment due to the advance of AI and robots. Hispanic and black workers are more at risk than white workers (47 percent and 44 percent versus 40 percent), and Asian workers are less at risk (39 percent).

Similar discriminations have been observed in other settings, such as advertising. Borgesius (2018) reports a study revealing that "when people searched for African-American-sounding names, Google displayed advertisements that suggested that somebody had an arrest record".[925] Needless to say, these associations might lead to discriminatory behaviour by decision-makers.

[919] Vassallo, T., E. Levy, M. Madansky, H. Mickell, B. Porter, M. Leas, and J. Oberweis, *Elephant in the Valley*, 2015. Retrieved at https://www.elephantinthevalley.com/.

[920] Todd, Sarah, "Inside the surprisingly sexist world of artificial intelligence," *Quartz*, October 25, 2015. Retrieved at https://qz.com/531257/.

[921] Leavy, Susan, "Gender bias in artificial intelligence: the need for diversity and gender theory in machine learning," *2018 ACM/IEEE 1st International Workshop on Gender Equality in Software Engineering*. DOI: 10.1145/3195570.3195580.

[922] Oudshoorn, Nelly, E. Rommes, and E. Stienstra, "Configuring the User as Everybody: Gender and Design Cultures in Information and Communication Technologies," *Science Technology Human Values*, Vol. 29, No. 1, 2004, pp. 30–63.

[923] Borgesius, Frederik Zuiderveen, *Discrimination, Artificial Intelligence and Algorithmic Decision-Making*, 2018. https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73

[924] Ibid.

[925] Ibid.

A more general treatment of AI and racial and ethnic bias in robots has been put forward by Sparrow (2019).[926] Sparrow argues that with regard to humanoid robots that operate in social settings, studies suggest that people are likely to attribute race to such systems. That is to say, if humanoid robots "have a race", it is likely that people view the races that they attribute to such robots as slaves. This is so, Sparrow argues, particularly if robots are perceived as being black, because of the historical background of slavery, and since the work done by robots is aimed to serve users and is based on a master-slave relationship. This, more generally, is an ethical problem since if robots are perceived both as humanlike and simultaneously as slaves, they might represent humans as slaves. Along these lines, Sparrow says that "the fact that humanoid robots refer to, and represent, human beings means that their design as machines intended to serve refers to the idea of human slaves".

A further issue that is connected to robots having a perceived race is the responsibility of engineers that design such systems. Even if they do not intend to design robots that people attribute race to, this nevertheless happens, and so raises the question as to how much engineers are responsible for how people interpret their design. An obvious solution to these issues is to design robots such that no race is attributed to them. Sparrow thinks that colouring robots blue or green, for example, might be a step in that direction.[927]

Empirically, it has been argued that attributing racial biases to robots might be due to social priming and moderated by the perceived anthropomorphism of such robots; doing away with what Sparrow suggests as a potential solution to the race attribution. Addison, Yogeeswaran, and Bartneck (2019) conducted two experiments based on the "shooter bias" paradigm to investigate the aforementioned phenomena of racial bias and perceived anthropomorphism.[928] The results showed that the shooter bias effect was still present for robots racialized as Black and White even in the absence of social priming. Interestingly, though, the study also revealed "that the shooter bias towards black robots disappeared when a brown robot was present no matter which robot type was encountered." By using differences in colours ranging from human to non-human like, the study aimed to find out "whether the shooter bias was influenced by how human-like the robot was." But this was not the case, since participants did not see the three differently coloured robots as differing in their perceived anthropomorphism. It should be noted, however, that much more work is needed to generalize the effects observed in these studies (having a less biased sample size, consisting not only of Caucasian participants, for one).

7.3.3. Children

In relation to children, the primary issue that is being discussed regarding AI and robots is how they affect children's cognitive, psychological and social development, and, following from this, how robots should be used in relation to children. Children are increasingly exposed to AI and robotic devices, for play, information, communication, education and therapy. As was discussed in section 7.1.11 on education and science, robots and AI program can have considerable educational benefits for children. However, there are some potential pitfalls as well. First, studies show that children are trusting towards AI programs and robots, and tend to believe what they say and have their opinions influenced,

[926] Sparrow, Robert, "Robotics Has a Race Problem," *Science, Technology, & Human Values*, 2019.
[927] Ibid.
[928] Addison, A., C. Bartneck, and K. Yogeeswaran, "Robots Can Be More Than Black And White," *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society - AIES 19*, 2019. doi: 10.1145/3306618.3314272

give in to social pressure exerted by these devices.[929,930] This introduces a serious risk of misuse, as well as harm because of erroneous performance by the device.

Having AI programs and robots as trusted sources, friends and role models means that such devices may transfer values, beliefs and viewpoints to children. A robot with opinions on what is right and wrong will influence a child's moral development. An AI program that is gendered or that voices explicit or implicit opinions on gender could shape children's views on and perceptions of gender. An AI program that makes everything into a competition teaches children to be competitive. Obviously, AI programs and robots are much more than pets or information sources, and their use with children introduces the need for strong protective measures.

Another concern is that children might mistake conversational AI programs and robots to be friends rather than pets or artefacts, and invest more in the relationship with them than in those with fellow children. Besides the loss of social interaction with real human beings, there is also the worry that such a development could threaten the development of empathy. Psychologist Sherry Turkle has argued that intelligent devices that present themselves as friends and objects worthy of empathy are deceitful and foster inauthentic empathy that does not involve the complexity and nuance involved in deep personal relationships.[931,932] The development of friendship bonds between robots and children has been a particular worry in the use or therapeutic robots for autistic children. Yet another worry is that AI and robots may in the future partially replace parents in the parenting role and drive a wedge between parents and children.[933]

Privacy is another concern. Intelligent devices typically collect vast amounts of information from their user in order to be able to interact successfully with them. Who has access to this information? This is especially a concern with internet-connected devices. The doll "My Friend Cayla", which is capable of real-time conversations with children, records the conversations and transmits them online to a voice analysis company.[934] Such a device is in breach of the UN Convention of the Rights of the Child and of the European General Data Protection Regulation. Yet, for other devices, it is not always clear what information they record and how it could be accessed to third parties.

7.3.4. The Elderly

In relation to the elderly, the primary issues besides privacy (as being discussed in the section on children), and data protection, are concerns whether AI and robots lead to isolation, loss of autonomy and dignity, and deception. We will turn to discussing the latter issues in some more detail, since they

[929] Vollmer, A., R. Read, D. Trippas, and T. Belpaeme, "Children conform, adults resist: A robot group induced peer pressure on normative social conformity," *Science Robotics*, Vol. 3, No. 2, 2018.

[930] Williams, Randi, Christian Vázquez Machado, Stefania Druga, Cynthia Breazeal, and Pattie Maes, ""My doll says it's ok": a study of children's conformity to a talking doll." In *Proceedings of the 17th ACM Conference on Interaction Design and Children* (IDC '18). ACM, New York, NY, USA, 625-631.

[931] Turkle, Sherry, "Authenticity in the age of digital companions," *Interaction Studies*, Vol. 8, No. 3, 2007, pp. 501–517.

[932] Turkle, Sherry, *Alone Together: Why We Expect More from Technology and Less from Each Other*, Basic Books, 2011.

[933] Havens, John, "Will we lose our rights as parents once robots are better at raising our kids?," *Quarts*, July 10, 2019. Retrieved at https://qz.com/co/2533915/.

[934] Firth-Butterfield, K., Generation AI: What happens when your child's friend is an AI toy that talks back? *World Economic Forum* website, 22 May 2018. Retrieved at https://www.weforum.org/agenda/2018/05/generation-ai-what-happens-when-your-childs-invisible-friend-is-an-ai-toy-that-talks-back/.

seem particularly pertinent to the elderly; whereas the former issues are more general concerns about AI and robots.

Regarding deception, there is a danger of users' inadequate expectations with regard to the functionality of robots that appear human-like. Akin to the danger of children viewing robots as friends, it has, for example, been observed that such robots induce the expectation in users of being able to converse in natural language with robots. If the robot is unable to do so, however, this can lead to frustration of users. Similar observations have been made regarding ascribing emotional states to robots, inducing the false expectation of the possibility to form emotional bonds between people and robots.[935] These ethical issues are connected to anthropomorphising such machines that can lead to inappropriate behaviour of users, such as creating a false sense of trust.

With regard to dignity, it has been argued that robots that aim to motivate the elderly to engage in conversations which raises potential problems of patronisation or infantilisation, as well as problems related to the aforementioned issues of making people believe they are interacting with a robot that they can potentially have a human-like relationship with. Communicating with robots has a related problem when it comes to social isolation. The more the elderly become capable of communicating with such machines, the more they might rely on such "relationships"; and conversely, the less they might feel the need for actual human conversations which might lead to social isolation.

When it comes to public opinion, Wachsmuth (2018) reported that out of 26000 European citizens completing a survey, "more than half (60%) of the respondents stated that the use of robots should be banned in the care of children, elderly, and the disabled."[936] Sparrow (2016) sees this worry based on two grounds: he thinks that robots are incapable of providing interpersonal relations of recognition and respect that are vital to promoting the well-being of the elderly.[937] Also, he thinks that it is likely that such systems will be used in institutional settings, and thus likely lead to replacing human caregivers that can provide interpersonal relations of recognition and respect. Thus, the overall level of care would be reduced.

It has also been argued that by introducing robots into the care of the elderly, the motivation might be more so to reduce costs and workload of human caregivers rather than improving the lives of the elderly. If robots are used to carry out highly personal tasks such as feeding, they might run the risk of making people feel "objectified," and thus reduce the level of well-being.[938]

Others argue that such dystopian scenarios are misleading since they fail to take into account that some elderly people may need care that does not treat them as (empirically) autonomous. It might also be that, in the future, the elderly are likely much more capable of using such systems properly than we imagine them now to be.[939]

The general point of contention can be regarded as different evaluations of the relationship between the elderly and robot caregivers. Opponents issue the worry that such relationships diminish the well-being of the elderly because they undermine values of respect, autonomy and dignity that are central

[935] Körtner, T., "Ethical challenges in the use of social service robots for elderly people," *Zeitschrift Für Gerontologie Und Geriatrie*, Vol. 49, No. 4, 2016, pp. 303–307.
[936] Wachsmuth, I., "Robots Like Me: Challenges and Ethical Issues in Aged Care," *Frontiers in Psychology*, Vol. 9, 2018. doi: 10.3389/fpsyg.2018.00432
[937] Sparrow, Robert, "Robots in aged care: a dystopian future?" *AI & Society*, Vol. 31, No. 4, 2016, pp. 445–454.
[938] Sharkey, Amanda, and Noel Sharkey, "Granny and the robots: ethical issues in robot care for the elderly," *Ethics and Information Technology*, Vol. 14, No. 1, 2010, pp. 27–40.
[939] Coeckelbergh, Mark, "Care robots and the future of ICT-mediated elderly care: a response to doom scenarios," *AI & Society*, Vol. 31, No. 4, 2015, pp. 455–462.

to human care. Whereas proponents are less sceptical, seeing more potential in furthering a fruitful application of such systems based on their efficiency.

7.3.5. People with physical and mental disabilities

In relation to people with physical and mental disabilities, the main issue that is being discussed regarding AI and robots can be summed up as the relation between newly gained opportunities of independence and increased risks of dependence.

It has been argued that an excessive use of technology to foster greater independence for people with disabilities could lead, unintendedly, to a new dependency on technology. By the same token, more opportunities to increase the autonomy of people with disabilities via AI and robotics might lead to the withdrawal of human caregivers, running a similar risk of social isolation that we have discussed in the previous section on the elderly.[940] That is to say, the worry is that once technological possibilities increase, perceived social responsibility of human caregivers might decrease to the disadvantage of people with disabilities.

Problems of social justice arise as well. How are we to distribute potentially expensive AI and robot systems to people in need? If only the affluent will benefit from such machines, we run the risk of further widening the gulf between rich and poor. It has been argued that the fairness issue at stake here is different for people with disabilities compared to attributes such as gender or race for two main reasons: on the one hand, there is an extreme diversity in the ways disabilities manifest, and people adapt. Also, since sharing disability information potentially leads to discrimination, it is not always disclosed.[941]

When it comes to the use of rehabilitation or therapy robots, it has been argued that people with mental disabilities might feel threatened by such devices, such that their usage might decrease their well-being instead of increasing it.[942]

The social pressure that rests on people with disabilities to make use of AI and robots once they are readily available might increase, since they might feel obligated to relieve human caregivers of their assistance even though they prefer human care over robotic care.

Regarding the use of robots for people with specific conditions such as autism, the worry has been raised that this might lead to understanding their condition as robotic like behaviour. Which can, in turn, be best treated with the aid of AI robotic assistance. If, for example, robots are used to teach people who are on the spectrum about social interaction, some experts think, this represents a severe misunderstanding of the condition.[943]

Notwithstanding the mentioned ethical concerns, it has been argued by Fiske et al. (2018) that, when implemented properly, AI and robots bear potential benefits for people with disabilities, such as expanding the reach of services to underserved populations or enhancing existing services provided

[940] Carnevale, A., "Robots, Disability, and Good Human Life", *Disability Studies Quarterly*, Vol. 35, No. 1, 2015.

[941] Trewin, S., "AI Fairness for People with Disabilities: Point of View," arXiv preprint arXiv:1811.10670, 2018.

[942] Tejima, N., "An ethical discussion on introducing rehabilitation robots for people with disabilities." *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*, 2009. doi: 10.1109/roman.2009.5326242

[943] Kobie, Nicole, "The questionable ethics of treating autistic children with robots," *WIRED*, July 18, 2018. https://www.wired.co.uk/article/autisim-children-treatment-robots

by mental health professionals.[944] Fiske et al., however, also think that AI and robots in the care of people with mental or physical disabilities should not be used to replace care by highly trained human health care professionals, since they can only ever be an assistance to traditional care. Supervised human care is also needed to minimize the risks associated with robotic care.

7.3.6. Educational and income level

In relation to educational and income level, the primary concerns regarding AI and robots are increased inequalities. Reasons for this include that the jobs that are potentially being replaced by AI and robots are more likely to be semi-administrative jobs that have previously been carried out by people with a lower level of education. These people, it has been argued, will have a more difficult time to find new jobs compared to people with higher educational whose social mobility might be less threatened due to their more transferable skills.[945]

When it comes to assessing the impact of AI and robots on income levels of the developed world, the aforementioned consequences might be bad but not fatal. Things could turn out differently regarding developing countries. The so-called "premature deindustrialization" might lead to a replacement of human labour with robots in countries that are not yet ready for that shift. Whereas it has been reported that in the US 47 percent of jobs are at risk of being replaced by AI and robots, in Ethiopia, for example, this figure is 85 percent.[946]

Schlogl and Sumner (2018) argue along those lines that the developing world might suffer more negative effects than the developed world for another reason besides labour substitution through AI and robots.[947] New industries, they say, may stop outsourcing production to the developing world since the work that previously has been done there at minimum wage, can now be done by robots here at even lower costs.

Some go as far to claim that the increase of AI and robots is directly linked to the increase of social inequality in more structural ways concerning education and income. The higher commerce sector, for example, continues to use human service providers, whereas the lower sector continuously replaces their service workers with AI and robots. It is likely, for example, that high-end stores will continue to provide human services to customers, whereas low-end stores will continue to lower costs by using AI and robots to serves their customers.[948]

[944] Fiske, A., P. Henningsen, and A. Buyx, "Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy (Preprint)" doi: 10.2196/preprints.13216, 2018.

[945] Vincent, James, "Robots and AI are going to make social inequality even worse, says new report," *The Verge*, July 13, 2017. https://www.theverge.com/2017/7/13/15963710/robots-ai-inequality-social-mobility-study

[946] Vincent, James, "First Click: Robots will make it even harder for poor countries to get rich," *The Verge*, March 10, 2016. https://www.theverge.com/2016/5/10/11648062/first-click-robots-will-make-it-even-harder-for-poor-countries-to-get

[947] Schlogl, L., and A. Sumner, A., "The Rise of the Robot Reserve Army: Automation and the Future of Economic Development, Work, and Wages in Developing Countries," *SSRN Electronic Journal*, 2018. doi: 10.2139/ssrn.3208816

[948] Marx, Paris, "Humans to serve the rich, robots to serve the poor," *Medium*, August 28, 2016. https://medium.com/radical-urbanist/humans-to-serve-the-rich-robots-to-serve-the-poor-6e2efc95c1b4

8. Conclusion

In this SIENNA deliverable, we have engaged in an extensive ethical analysis of artificial intelligence and robotics technologies, including their various manifestations and applications. Its primary aims have been to identify and analyse ethical issues in AI and robotics, both present and potential future ones, with a time horizon of twenty years. We have not tried to make recommendations or present solutions, but only to identify and analyse ethical issues. A secondary aim of this report has been to convey the results of SIENNA's "country studies" of the national academic and popular media debate on the ethical issues in AI and robotics in twelve different EU and non-EU countries, highlighting the similarities and differences between these countries. In what follows, we provide a summary of our findings and give a brief outline of how this report will be used for further work in the context of SIENNA.

To begin with, our analysis of the "country studies" (in section 4 of this report) that were carried out in eight EU and four non-EU countries produced a number of interesting findings. Our analysis of the *national academic debates* found that across all twelve countries, the most widely discussed application areas of AI and robotics are defence, medicine, transportation, and the workplace, with the most-discussed products being autonomous weapon systems (especially "killer robots"), care robots, healthcare apps, surgical robots, sex robots, and autonomous vehicles. Especially notable was the significant amount of attention the ethics of defence applications of AI and robotics in most countries. In most countries, a wide range of ethical issues were discussed, relating to justice, equality, autonomy, dignity, explainability, transparency, safety, accountability, liability, privacy, and data protection. This largely reflects the international academic debate. The most frequently mentioned issues were justice, privacy, and safety, which were often still addressed in countries were academic discussion was found to be scant. The national academic debates in the US, Germany and China stood out in also being focused on potential broad-scoped solutions to ethical issues, including through laws, standards, and regulation, as well as through ethics by design and implementation of moral reasoning systems in robots and AI systems.

In our study of national *popular media debates*, we observed that in all countries, with the possible exception of Poland, there has been substantial debate in the national popular media on ethical issues in relation to AI and robotics, although in some countries the debate has only recently gained pace. In most cases, the application areas, products, and ethical issues and principles addressed in the popular academic debate mirrored those in the academic debate. Issues related to the potential economic effects of AI and robotics technology, however, seemed to get slightly more attention.

After presenting SIENNA's "country studies" results, this deliverable reported on the broad-scoped ethical analysis (in sections 5, 6 and 7) that was conducted using the SIENNA approach to ethical analysis (which was presented in section 2), and which featured extensive literature review, consultation of experts and stakeholders, and original ethical analysis. This analysis had three parts, in which we discussed the following.

In the first part of our ethical analysis of AI and robotics (in section 5), we covered general ethical issues with AI technology and robotics technology: issues with the aims of these technologies, issues with their techniques and approaches, and issues in terms of their risks and implications. We first analysed ethical issues associated with the general aims of AI and robotics technology. It was found, amongst others, that the aims of efficiency, productivity and effectiveness improvement through AI and robotics are inherently tied to the replacement of human workers, which raises ethical issues. We also found that the aim of mimicking of social behaviour in AI and robotics is associated with risks of deception and of diminished human-to-human social interaction. Further, we found that the aim of developing artificial general intelligence and superintelligence raises issues of human obsolescence and loss of control, and raises issues of AI and robot rights. The aim of human cognitive enhancement, finally, was found to bring risks to equality, human psychology and identity, human dignity and privacy.

Next, we discussed ethical issues associated with techniques and approaches in AI and robotics. For AI, these issues included the following. In relation to algorithms, we discussed how they can be value-laden and contain biases. In relation to knowledge representation, we discussed how inaccuracy, misrepresentation and bias can raise ethical issues. We discussed how automated scheduling and planning can raise issues of trustworthiness and responsibility, and could decrease human capabilities. In relation to machine learning, we discussed many ethical issues, including issues of transparency and explainability, fairness and discrimination, reliability, privacy and accountability. Machine ethics was analysed to have many pitfalls, including the difficulty of implementing human morality in AI systems, the potential for failure and corruptibility, equality of access to ethical AI, the undermining of human moral responsibility, and the possibility that we want to grant such systems moral status and rights.

The issues with robotics techniques and approaches were the following. For robot sensing, issues of reliability of error were discussed, as well as risks to privacy and safety associated with some sensor types. In relation to robot actuation, we discussed issues of safety, privacy, and psychological impacts. And for robot control systems, we discussed how robots can have different degrees of autonomy, and we discussed associated issues of safety, responsibility and accountability, transparency, and privacy.

Finally, we described a number of general implications and risks associated with the development and use of AI and robotics. For AI, these included potential negative implications for autonomy and liberty, privacy, justice and fairness, responsibility and accountability, safety and security, dual use and misuse, mass unemployment, transparency and explainability, meaningfulness, democracy and trust. For each value or issue, we aimed to come to a precise determination of it, we then discussed different general ways in which AI might impact it, and we analysed the moral considerations involved. For robotics, the general implications and risks included loss of control, autonomy, privacy, safety and security, dual use and misuse, mass unemployment, human obsolescence, human mistreatment, robot rights, and responsibility and accountability. We analysed these issues like we did in the corresponding part on AI.

In the second part of our ethical analysis of AI and robotics (in section 6), we covered ethical issues with specific products, systems and processes in AI and robotics. For AI, these issues included the following. In relation to intelligent agents, we found ethical issues that include privacy, user autonomy and authentic personhood, trust, moral responsibility and liability, and questions about how ethical behaviour is best instilled in these constructs. With respect to knowledge-based systems, we identified issues that include bias in knowledge representation and inferential patterns, self-modification of such systems that leads to unpredictable outcomes, accuracy, and security. In relation to computer vision systems, we found ethical concerns in relation to object detection, image classification, object recognition, and visual biometric applications, involving security, accuracy, and privacy. Natural language processing systems were found to raise issues of privacy, and potential bias and discrimination in algorithms and use of data. For affective computing systems, issues were identified that involved privacy and trust, as well issues with using affective capabilities for deception, and unwanted social bonding and loss of autonomy. In relation to (big) data analytics systems, major issues of individual and group privacy, potential algorithmic bias and discrimination, and issues of transparency and accountability were identified. With respect to embedded AI & Internet-of-Things, finally, we analysed concerns about the implications of their use in terms of privacy, security and trust, autonomy and freedom, and accountability.

Various important types of robotic systems raised the following important or unique ethical concerns. We found that humanoid robots could easily become the subject of misplaced moral accountability, misplaced trust, and misplaced empathy, and could reinforce stereotypes and be used to perpetuate socially undesirable behaviour. Social robots, were found to raise many of the same concerns as humanoid robots, and also raise the broader question about the (social) contexts in which they should or should not be used. For unmanned aerial vehicles, or drones, we identified issues of privacy, accountability, security, and transparency, and more generally the uses to which they should be put. In relation to autonomous vehicles, we found issues of privacy, accountability, security and transparency, as well as issues concerning the implemented crash algorithms, and the way in which autonomous vehicles make decisions in general. For telerobotic systems, we identified issues in terms of diminished social interaction between humans, negative effects on the psychological well-being of operators, and specific harms from increased technologisation, as well as issues of safety, security, equality, and responsibility. Robotic exoskeletons were found to raise issues of possible negative physical and psychological impacts on users, issues of access and equality, privacy, safety, and security, and the possibility of dehumanization or overworking of industrial labourers. For biohybrid robots, we identified issues concerning their moral status and permissibility. In relation to swarm robots, we found that they raise concerns because of their great potential for surveillance, and their potential unpredictability and uncontrollability, and that safety, security and dual-use are also concerns. For microrobots, we identified issues of surveillance and privacy, control and ownership, safety, and environmental degradation. Collaborative robots, finally, were found to raise issues of trust and risks of psychological harm for human co-workers, and issues of privacy and security.

In this third part of our ethical analysis (in section 7), we covered ethical issues with the application of AI and robotics in different application domains, and ethical issues for different types of users and stakeholders. For AI applications, we identified the following major application domains: infrastructure and cities, healthcare, finance and insurance, defence, law enforcement, the legal sector, public services and governance, retail and marketing, media and entertainment, smart home and companionship, education and science, manufacturing, and agriculture. Recurring ethical issues in these different domains were found to include privacy, transparency, responsibility, fairness, freedom, autonomy, security and trust. For domains in which they are an issue, we discuss their particular manifestations and peculiarities.

Healthcare applications of AI were found to raise special issues regarding potential risks to privacy and trust, threats to informed consent, discrimination, and risks of further increasing already existing health inequalities. For law enforcement applications, we identified issues of bias and discrimination, surveillance, and the risk of a lack of accountability and transparency for law enforcement decisions. It was found that defence applications come with possible negative effects of AI on compliance with the principles of just war and the law of armed conflict, the possibility for uncontrolled or inexplicable escalation, and the potential for responsibility gaps. In media and entertainment, we discussed ethical issues in news media, social media and audio and visual media. In news media, there is the risk of impoverished journalism, hyper-personalization that contributes to “filter bubbles”, and smart generation of fake news. In audio and visual media, like film and music, we found that AI could undermine creativity if pushed too far, instituting formulaic processes that lack the creativity, spontaneity and humanity that human creators can bring. For social media, we determined that harvesting of personal information for advertising and political microtargeting could undermine privacy and democracy, that AI could stimulate the formation of “echo chambers”, and that there are controversies around automated social media censorship. Finally, we found that AI in the agricultural sector could further increase the power imbalance between agribusinesses and farmers, and could reinforcing big industrial monocultures. Other application domains were also found to raise various unique issues.

For robotics applications, we identified the following application domains that raise important or unique ethical concerns: transportation, law enforcement, defence, infrastructure, healthcare, companionship, manufacturing, exploration, service sector, and environment and agriculture. Frequently recurring ethical issues in these domains were found to include privacy, transparency, responsibility, fairness, autonomy, safety and trust. For domains in which they are an issue, we discussed their particular manifestations and peculiarities.

We found that transportation applications, involving automated vehicles, raise significant issues, of trust, accountability, transparency, security and safety. In healthcare, we found issues of patient privacy and confidentiality, maintenance of quality of care and patient integrity, and the risks of reduced humanity in patient care. The area of companionship was found to include ethical issues involving security, privacy and safety, possible negative implications for human-human interaction, and the appropriate of certain applications of companion robots, for example for child care, elderly care, and sex and romantic relationships. In the service sector, including retail, recreation, restaurants, banking, and communications, amongst others, one issue was found regarding the extent to which robots should be able to make decisions without human approval or interference, and the value trade-offs this involves. Two other issues concerned the replacement of human workers by service robots, and the risk of resemblances to slavery in certain service robot applications. The other mentioned application domains were also found to raise various special ethical issues.

Finally, we identified and described the following ethical issues that concern different types of end users and other stakeholders of AI and robotics technologies. With respect to gender, ethical issues include the possibility of women being disproportionally affected by AI-induced unemployment, algorithmic and functional gender bias and gender stereotyping in the design of AI and robotics products, and the lack of women in the AI and robotics technology sectors. With regard to race and ethnicity, ethical issues include algorithmic racial bias in the design of AI products, and humanoid robots contributing to the perception of particular racial groups in society as slaves. With respect to children, ethical issues include the shaping of children's views by biased AI systems and robots, a potential loss of social interaction with other children, stunted empathy development in children, and potential harms to privacy by intelligent Internet-connected toys. With regard to the elderly, ethical issues include potential harms to privacy, the generation of false expectations about the (social) abilities of anthropomorphic robots, the potential for patronisation of elderly individuals by robots, and a potential loss of social interaction with other human beings. With regard to people with physical and mental disabilities, ethical issues include risks of dependency on AI systems and robots and increased social isolation, a diminished perception of social responsibility among human caregivers, and distributive justice concerns. With respect to educational and income level, ethical issues include unequal effects of AI and robotics on people depending on their level of education, and increased inequalities between the developed world and the developing world.

Having now summarised the most important findings of this deliverable, let us conclude by briefly looking at further work in the context of SIENNA. As stated earlier, the aim of the report has not been to make recommendations or present solutions, but only to identify and analyse ethical issues. The report charts the ethical issues that should be taken into account in the development, use and regulation of AI and robotics technologies along their full breadth. In SIENNA, the findings presented here provide an important basis for our next report (SIENNA D4.7, which is due in 2020), in which we aim to present an ethical framework for AI and robotics that contains recommendations and solutions for ethical issues. This will bring us one step closer to realising the project's aims of developing a set of practical tools including new operational guidelines for research ethics committees, codes of responsible conduct and policy recommendations, which we hope will contribute to a responsible future development and use of AI and robotics technologies.

All deliverables of the SIENNA project can be found on its website, at the following address: http://www.sienna-project.eu/publications/deliverable-reports/.

9. References

Aaron Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science,* Vol. 9, No. 3–4, 2014, pp. 211-407.

Abbasi, Sarker, and Chiang, “Big Data Research in Information Systems”.

Abouelmehdi, Karim, et al., “Big Data Security and Privacy in Healthcare: A Review,” *Procedia Computer Science*, Vol. 113, 2017.

Acemoglu, Daron, “Why Universal Basic Income Is a Bad Idea,” *Project Syndicate*, June 7, 2019. https://www.project-syndicate.org/commentary/why-universal-basic-income-is-a-bad-idea-by-daron-acemoglu-2019-06

Acemoglu, Daron, and Restrepo, Pascual, “Artificial Intelligence, Automation and Work”, National Bureau of Economic Research (NBER), January 2018, p. 4.

Ackerman, Evan, “Robotic Tortoise Helps Kids to Learn That Robot Abuse is a Bad Thing”, *Spectrum IEEE*, March 14.

Adam, Alison, *Artificial Knowing: Gender and the Thinking Machine*, Florence, KY, USA: Routledge, 1998.

Addison, A., C. Bartneck, and K. Yogeeswaran, “Robots Can Be More Than Black And White,” *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society - AIES 19*, 2019. doi: 10.1145/3306618.3314272

Aetheon. “Tug Informational Graphics”, 2018. Accessed December 2018. aethon.com/infographics

Ahmed, Kaoutar Ben, Mohammed Bouhorma, and Mohamed Ben Ahmed, "Age of big data and smart cities: privacy trade-off," *arXiv preprint arXiv:1411.0087*, 2014.

Akhavan et al., “Exploring the Relationship between Ethics, Knowledge Creation and Organizational Performance”, 44.

Akhil, A., “Researchers Aim to Build Eco-Friendly Robots with Biodegradable Materials”, *Sastra Robotics*, July 2018.

Al-Naji, Ali, Perera, Asanka & Chahl, Javaan, “Remote Monitoring of Cardiorespiratory Signals from a Hovering Unmanned Aerial Vehicle”, *BioMedical Engineering Online* 16(101), August 2017.

AlDairi, Anwaar, and Lo’ai Tawalbeh, "Cyber security attacks on smart cities and associated mobile technologies," *Procedia Computer Science,* Vol. 109, 2017, pp. 1086-1091.

Alesich, Simone, and Michael Rigby, “Gendered Robots: Implications for Our Humanoid Future,” *IEEE Technology and Society Magazine*, Vol. 36, No. 2, 2017, pp. 50-59.

Alexander R. Bentley, O'Brien, M., J. & Brock, W. A., (2014). Mapping collective behavior in the big-data era. Behavioral and Brain Sciences 37 (1):63-76.

Allen, Colin, and Wendell Wallach, *Moral Machines: Teaching Robots Right from Wrong*, London, U.K.: Oxford University Press, 2009.

Allen, Colin, Gary Varner, and Jason Zinser, “Prolegomena to Any Future Artificial Moral Agent,” *Journal of Experimental & Theoretical Artificial Intelligence*, Vol. 12, No. 3, 2000, pp. 251–261.

Allen, Colin, Wendell Wallach, and Iva Smit, “Why Machine Ethics?,” *IEEE Intelligent Systems*, Vol. 21, No. 4, 2006, pp. 12–17.

Amini, Alexander, Ava Soleimany, Wilko Schwarting, Sangeeta Bhatia, and Daniela Rus, "Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure," *Proceedings of the 2019 AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES), 27-28 January, 2019 Honolulu, Hawaii, United States, AAAI/ACM,* 2019.)

Amiribesheli, Mohsen, Asma Benmansour, and Abdelhamid Bouchachia, "A review of smart homes in healthcare," *Journal of Ambient Intelligence and Humanized Computing,* Vol. 6, No. 4, 2015, pp. 495-517., p. 495

Amit Datta, Tschantz M. C. & Datta, A. (2015). Automated Experiments on Ad Privacy Settings A Tale of Opacity, Choice, and Discrimination. Proceedings on Privacy Enhancing Technologies 2015 (1), 92–112.

Amnesty International, "Autonomous Weapons Systems: Five Key Human Rights Issues for Consideration", 2015. https://www.amnesty.org/en/documents/act30/1401/2015/en/

Ananny, Mike, and Kate Crawford, "Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability," *New Media & Society*, Vol. 20, No. 3, 2016, pp. 973–989.

Anderson and Anderson, "Toward Ensuring Ethical Behavior from Autonomous Systems", 2.

Anderson, Susan Leigh, "Machine metaethics," in *Machine Ethics*, M. Anderson and S. Anderson, Eds., New York, NY, USA: Cambridge University Press, 2011, pp. 21–27.

André, Quentin, et al., "Consumer Choice and Autonomy in the Age of Artificial Intelligence and Big Data," *Customer Needs and Solutions*, Vol. 5, p. 28–37.

Andreas Matthias (2004) The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology 6(3): 175–183

Andrew D. Selbst, 'Disparate Impact in Big Data Policing', *Georgia Law Review*, Vol. 52, No. 1, 2017, p. 109. Concept Paper of the 2019 OSCE Annual Police Experts Meeting *Artificial Intelligence and Law Enforcement: An Ally or an Adversary?,* 23-24 September 2019, Vienna: https://polis.osce.org/2019APEM

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

Antoinette Rouvroy; Rouvroy, Antoinette, and Berns, Thomas, "Algorithmic Governmentality and Prospects of Emancipation," *Réseaux*, Vol. 177, No. 1, 2013.

Aparna, Kale. Bodhale, Umesh "Overview Of Sensors For Robotics", International Journal of Engineering Research and Technology (IJERT) Volume 02, Issue 03 (March 2013).

Applin, Sally, "Autonomous Vehicles Ethics, Stock or Custom?", *IEEE Consumer Electronics* 6(3), June 2017.

Araya, Agustin A., "Questioning Ubiquitous Computing," *Proceedings of the 1995 ACM 23rd Annual Conference on Computer Science - CSC 95*, 1995.

Arkin, Ronald, *Governing Lethal Behavior in Autonomous Robots*. Boca Raton, Chapman and Hall/CRC Press, 2009.

Arnold, Thomas & Scheutz, Matthais, "The Tactile Ethics of Soft Robotics: Designing Wisely for Human-Robot Interaction", *Soft Robotics* 4(2), 2017.

Arruda, Andrew, "An Ethical Obligation to Use Artificial Intelligence: An Examination of the Use of Artificial Intelligence in Law and the Model Rules of Professional Responsibility," *American Journal of Trial Advocacy*, Vol. 40, 2017, pp. 443–58.

Arruda, Andrew, "The world's first AI legal assistant", TED Talk, November 2018. https://www.ted.com/talks/andrew_arruda_the_world_s_first_ai_legal_assistant

Asaro, Peter M., 'AI Ethics in Predictive Policing. From Models of Threat to an Ethics of Care', *IEEE Technology and Society Magazine*, June 2019, pp. 44–46.

Asaro, Peter, "Algorithms of Violence: Critical Social Perspectives on Autonomous Weapons", *Social Research*, Vol. 86, No. 2, 2019, p. 539.

Asaro, Peter, "On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making", *International Review of the Red Cross*, Vo. 94, 2012, pp 687-709.

Asaro, Peter, "What Is an Artificial Intelligence Arms Race Anyway", *I/S: Journal of Law and Policy for the Information Society*, Vol. 15, 2019, pp. 45-64.

Asaro, Peter, “Why the world needs to regulate autonomous weapons, and soon”, *Bulletin of the Atomic Scientists*, 27 April 2018. https://thebulletin.org/2018/04/why-the-world-needs-to-regulate-autonomous-weapons-and-soon/

Asaro, Peter. W., “The labor of surveillance and bureaucratized killing: new subjectivities of military drone operators”, *Social semiotics*, Vol. *23*, No. 2, 2013, p. 220.

Athalye, Anish, Logan Engstrom, Andrew Ilyas, and Kevin Kwok, “Synthesizing Robust Adversarial Examples,” Cornell University arXiv.org, 2018. https://arxiv.org/abs/1707.07397

Avgousti, Sotiri, et al., “Medical telerobotic systems: current status and future trends”, *Biomedical Engineering OnLine*, Vol. 15, No. 96, 2016.

Ayadi, Moataz El, Mohamed S. Kamel, and Fakhri Karray, “Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases,” *Pattern Recognition,* Vol. 44, No. 3, 2011, pp. 572–587.

Bali, Meghna, “Companion Robots: What are the Ethical Implications of Intimate Human-Machine Relationships?”, *ABC News*, August 2017.

Barlow, Rich, “Economist Predicts Job Loss to Machines, but Sees Long-Term Hope”, *Psys.org Robotics*, March 2018.

Barocas, Solon, and Andrew D. Selbst, “Big Datas Disparate Impact,” *Calif. L. Rev.,* Vol. 104, 2016, p. 671.; Brynjolfsson, Erik, and Tom Mitchell, “What Can Machine Learning Do? Workforce Implications,” *Science,* Vol. 358, No. 6370, 2017, pp. 1530–1534.

Basu, Bhattacharyya, and Kim, “Use of Artificial Neural Network in Pattern Recognition”, 24.

Batchelor, Bruce G. & Whelan Paul F., “Ethical, Environmental and Social Issues for Machine Vision in Manufacturing Industry”, *Machine Vision Applications, Architectures, and Systems Integration IV*, 1995.

Beam, Andrew L. and Kohane, Isaac S., “Big Data and Machine Learning in Health Care,” *Journal of American Medical Association*, March 2018, E1–2.

Beard, Matthew, “With robots, is a life without work one we'd want to live?,” *The* Guardian, September 26, 2016. https://www.theguardian.com/sustainable-business/2016/sep/26/with-robots-is-a-life-without-work-one-wed-want-to-live

Beatriz Cardona (2008) ‘Healthy ageing’ policies and anti-ageing ideologies and practices: On the exercise of responsibility. Medicine, Health Care and Philosophy 11(4): 475–483

Beauchamp, Tom L., and Childress, James F., *Principles of Biomedical Ethics*, Oxford, Oxford University Press, 2012. We are grateful to Tally Hatzakis for reviewing this section.

Bekey, George A., *Autonomous Robots: From Biological Inspiration to Implementation and Control*, MIT Press, February 2017.

Bendel, Oliver, “Co-Robots from an Ethical Perspective”, *Business Information Systems and Technology 4.0*, March 2018.

Berghel, H., ‘Malice Domestic: The Cambridge Analytica Dystopia’, *Computer*, Vol. 51, No. 5, pp. 84-89, May 2018.

Berlin, Isaiah, Two Concepts of Liberty, 1969, in Berlin, Isaiah, *Four Essays on Liberty*, Oxford University Press, Oxford, p. 118-172.

Bernard, Pascal, “Is AI a threat to Democracy?,” *Towards data Science*, May 21, 2019. https://towardsdatascience.com/is-ai-a-threat-to-democracy-4bef3e5fcfdd

Bhoge, Anand, “Smart Robotics: Revolution is Motto, Efficiency is Aim”, *Robotics Tomorrow*, August 2018.

Bigda, Jordan, “The Legal Profession: From Humans to Robots,” *Journal of High Technology Law*, Vol. 18, 2018, pp. 396–428. Seedrs in “Six ways the legal sector is using AI right now” (13 December 2018) identifies six different aspects of this first type of use of AI in law. https://www.lawsociety.org.uk/news/stories/six-ways-the-legal-sector-is-using-ai

Binns, R. (2017). Algorithmic Accountability and Public Reason. *Philosophy and Technology*, 1-14.

Bisconti Lucidi, Piercosma & Nardi, Daniele, “Companion Robots: The Hallucinatory Danger of Human-Robot Interactions”, *Association for the Advancement of Artificial Intelligence*, 2018.

Bissolotti, Luciano, Nicoli, Federico & Picozzi, Mario, “Domestic Use of the Exoskeleton for Gait Training in Patients with Spinal Cord Injuries: Ethical Dilemmas in Clinical Practice”, *Frontiers in Neuroscience*, February 2018.

Blodgett, Su Lin, Lisa Green, and Brendan O’Connor, “Demographic Dialectal Variation in Social Media: A Case Study of African-American English,” *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.

Blodgett, Su Lin, Lisa Green, and Brendan O’Connor, “Demographic Dialectal Variation in Social Media: A Case Study of African-American English,” *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016., p. 1

Bohn, J., V. Coroamă, M. Langheinrich, F. Mattern, and M. Rohs, “Social, Economic, and Ethical Implications of Ambient Intelligence and Ubiquitous Computing,” *Ambient Intelligence*, 2005, pp. 5–29.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A., ‘Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings’, 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, 2016. http://papers.nips.cc/paper/6227-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," In *Advances in neural information processing systems*, pp. 4349-4357, 2016.

Bonnefon, Jean-François, Shariff, Azim & Rahwan, Iyad, “Autonomous Vehicles Need Experimental Ethics: Are We Ready for Utilitarian Cars?” 2015.

Booth, Serena, Tompkin, James & Pfister, Hanspeter et al., “Piggybacking Robots: Human-Robot Overtrust in University Dormitory Security”, *Human Robot Interaction*, March 2017.

Borgesius, Frederik Zuiderveen, *Discrimination, Artificial Intelligence and Algorithmic Decision-Making*, 2018. https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73

Borzaga, Carlo, Gianluca Salvatori, and Riccardo Bodini, “Social and Solidarity Economy and the Future of Work,” *Journal of Entrepreneurship and Innovation in Emerging Economies,* Vol. 5, No. 1, 2019, pp. 37-57. https://www.ilo.org/global/topics/cooperatives/publications/WCMS_573160/lang--en/index.htm.

Bostrom, Nick, “Ethical Issues in Advanced Artificial Intelligence,” 2003. https://nickbostrom.com/ethics/ai.html.

Bozdag, Engin, and Jeroen Van Den Hoven, “Breaking the Filter Bubble: Democracy and Design,” *Ethics and Information Technology,* Vol. 17, No. 4, 2015, pp. 249–265.

Braun, Ashley. “The RangerBot is a New Line of Defense Against Coral-Eating Crown-of-Thorns Starfish”, Smithsonian, August 2018.

Braun, Trevor, Benjamin CM Fung, Farkhund Iqbal, and Babar Shah. "Security and privacy challenges in smart cities," *Sustainable cities and society,* Vol. 39, 2018, pp. 499-507., p. 2

Brayne, Sarah, “Big Data Surveillance: The Case of Policing,” *American Sociological Review*, Vol 82, No. 5, 2017.

Breazeal, Cynthia, “Toward sociable robots,” *Robotics and Autonomous Systems*, Vol. 42, 2003, pp. 167–175.

Brey, P. (2010). Values in Technology and Disclosive Computer Ethics. In L. Floridi (Ed.), *The Cambridge Handbook of Information and Computer Ethics* (pp. 41-58). Cambridge: Cambridge University Press.

Brey, P.A.E., “Anticipatory Ethics for Emerging Technologies”, *Nanoethics*, Vol. 6, 2012, pp. 1–13.

Brey, Philip, “Freedom and Privacy in Ambient Intelligence,” *Ethics and Information Technology,* Vol. 7, No. 3, 2005, pp. 157–166., p. 8.

Brey, Philip, “Freedom and Privacy in Ambient Intelligence,” *Ethics and Information Technology,* Vol. 7, No. 3, 2005, pp. 157–166., p. 9.

Brinton, Chris, "A framework for explanation of machine learning decisions," In *IJCAI-17 Workshop on Explainable AI (XAI)*, 2017, pp. 14-18., p. 14

Bronson, Kelly, and Knezevic, Irenam “Big Data in Food and Agriculture,” *Big Data & Society*, 2018, p. 2.

Brooks, Victoria, “Samantha’s Suffering: Why Sex Machines Should Have Rights Too”, *The Conversation*, April 2018.

Brown, Robert, “Robots Make A Money-Making Assembly Line by Cutting Costs”, *Center for The Future of Work*, February 2015.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., Ó hÉigeartaigh, S., Beard, S., Belfield, H., Farquhar, S., Lyle, C., Crootof, R., Evans, O., Page, M., Bryson, J., Yampolskiy, R., and Amodei, D., ‘The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation’, 2018. https://www.eff.org/files/2018/02/20/malicious_ai_report_final.pdf.

Brundage, Miles, “Limitations and risks of machine ethics,” *Journal of Experimental & Theoretical Artificial Intelligence*, Vol. 26, No. 3, 2014, pp. 355–372.

Bryant, “Knowledge Management — The Ethics of the Agora or the Mechanisms of the Market?”

Brynjolfsson, Erik, and Tom Mitchell, “What Can Machine Learning Do? Workforce Implications,” *Science*, American Association for the Advancement of Science, December 22, 2017. http://science.sciencemag.org/content/358/6370/1530.

Brynjolfsson, Erik, and Tom Mitchell, “What Can Machine Learning Do? Workforce Implications,” *Science,* Vol. 358, No. 6370, 2017, pp. 1530–1534.

Brynjolfsson, Erik, and Tom Mitchell, “What Can Machine Learning Do? Workforce Implications,” *Science,* Vol. 358, No. 6370, 2017, pp. 1530–1534.; Litvinski, O. (2018). Algorithmic opacity: a narrative revue.

Bryson, J. J. (2010). Robots should be slaves. In Y. Wilks (Ed.), Close engagements with artificial companions: Key social, psychological, ethical and design issues (pp. 63–74). Amsterdam: John Benjamins.

Bullington, Joseph, “Affective computing and emotion recognition systems: The future of biometric surveillance?,” *Information Security Curriculum Development Conference ‘05*, 2005.

Bundy, Alan, and Fiona Mcneill, “Representation as a Fluent: An AI Challenge for the Next Half Century,” *IEEE Intelligent Systems*, Vol. 21, No. 3, 2006, pp. 85–87. https://ieeexplore.ieee.org/abstract/document/1637360

Buolamwini, J., and Gebru, T., ‘Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification’, Proceedings of the 1st Conference on Fairness, Accountability, and Transparency, PMLR, Vol. 81, pp. 77-91, 2018.

Buoncompagni, Luca, Capitanelli, Alessio, & Carfi, Alessandro et al., “From Collaborative Robots to Work Mates: A New Perspective on Human-Robot Cooperation”, *ERCIM News*, July 2018.

Burgess, Matt, “Now DeepMind's AI can spot eye disease just as well as your doctor”, *Wired*, August 13, 2018. https://www.wired.co.uk/article/deepmind-moorfields-ai-eye-nhs

Burke, “Knowledge-Based Recommender Systems”.

Burr, Chistopher, Nello Cristianini, and James Ladyman, “An Analysis of the Interaction Between Intelligent Software Agents and Human Users,” *Minds and Machines*, Vol. 28, No. 4, 2018, pp. 735–774.

Burrell, J., ‘How the Machine ‘Thinks’: Understanding Opacity in Machin Learning Algorithms’*, Big Data & Society*, Vol. 3, No. 1, June 2016.

Burrell, Jenna, “How the Machine Thinks: Understanding Opacity in Machine Learning Algorithms,” *Big Data & Society,* Vol. 3, No. 1, 2015, p. 2053951715622512.

Cammozzo, A., ‘Face Recognition and Privacy Enhancing Techniques’, in Bissett, A., Bynum, T. W., Light, A., Lauener, A., and Rogerson, S., ETHICOMP 2011: The Social Impact of Social Computing, Sheffield, UK, Sheffield Hallam University, pp. 101- 109, 2011.

Cangelosi, Angelo & Schlesinger, Matthew, “From Babies to Robots: The Contribution of Developmental Robotics to Developmental Psychology”, *Child Development Perspectives,* February 2018.

Capgemini, “Microbots: Innovation in Healthcare”, 2 Dec 2014. https://www.capgemini.com/2014/12/microbots-innovation-in-healthcare-0/

Carbonell, Isabelle M., “The Ethics of Big Data in Big Agriculture,” *Internet Policy Review*, Vol. 5, No. 1, 2016, p. 3.

Carlini, Nicholas, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou, "Hidden voice commands," *25th {USENIX} Security Symposium ({USENIX} Security 16)*, pp. 513-530, 2016., p. 513

Carlisle, Brian, “Pick and Place for Profit: Using Robot Labor to Save Money”, *Robotics Business Review*, September 2017.

Carnevale, A., “Robots, Disability, and Good Human Life”, *Disability Studies Quarterly*, Vol. 35, No. 1, 2015.

Carolan 2015 quoted in Mark Ryan, “Ethics of Using AI and Big Data in Agriculture: The Case of a Large Agriculture Multinational,” *ORBIT Journal*, Vol. 2, No. 2, 2019, p. 6. https://doi.org/10.29297/orbit.v2i2.109

Castelvecchi, D., ‘The Black Box of AI’, *Nature*, Vol. 538, No. 7623, pp. 20-23, 6 October 2016.

Cave, Stephen, Rune Nyrup, Karina Vold, and Adrian Weller, “Motivations and Risks of Machine Ethics,” *Proceedings of the IEEE*, Vol. 107, No. 3, 2019, pp. 562–574.

Char, Danton S., Shah, Nigam H., and Magnus, David, “Implementing Machine Learning in Health Care — Addressing Ethical Challenges,” *New England Journal of Medicine*, Vol. 378, No. 11, March 2018, p. 3.

Charisi, Vicky, Louise Dennis, Michael Fisher, Robert Lieck, Andreas Matthias, Marija Slavkovik, Janina Loh (Sombetzki), Alan F.T. Winfield, and Roman Yampolskiy, “Towards moral autonomous systems.” Cornell University arXiv.org, 2017. https://arxiv.org/abs/1703.04741

Chatila, Raja, “Inclusion of Humanoid Robots in Human Society: Ethical Issues”, *Humanoid Robots: A Reference*, October 2017.

Chesney, B., and Citron, D., ‘Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security’, SSRN, 2018. https://ssrn.com/abstract=3213954

Choudhary, Harding, and Tiwari, “Data Mining in Manufacturing”.

Christman, John, “Autonomy in Moral and Political Philosophy,” *Stanford Encyclopedia of Philosophy*, Edward N. Zalta, January 9, 2015. https://plato.stanford.edu/entries/autonomy-moral/

Chung, Hyunji, Michaela Iorga, Jeffrey Voas, and Sangjin Lee, "Alexa, can I trust you?," *Computer,* Vol 50, no. 9, 2017, pp. 100-104.

Civil Rights Groups, *Predictive Policing Today: A Shared Statement of Civil Rights Concerns*, 31 August 2016 <https://www.aclu.org/other/statement-concern-about-predictive-policing-aclu-and-16-civil-rights-privacy-racial-justice> [accessed 1 July 2019].

Clarke, Roger, *Introduction to Dataveillance and Information Privacy, and Definitions of Terms*, 1997. http://www.rogerclarke.com/DV/Intro.htm

Coeckelbergh, Mark, “Care robots and the future of ICT-mediated elderly care: a response to doom scenarios,” *AI & Society*, Vol. 31, No. 4, 2015, pp. 455–462.

Coeckelbergh, Mark, “Health Care, Capabilities, and AI Assistive Technologies,” *Ethical Theory and Moral Practice*, Vol. 13, 2010.

Cohen, Noam, “Will California’s New Bot Law Strengthen Democracy?,” *The New Yorker*, 2 July 2019. https://www.newyorker.com/tech/annals-of-technology/will-californias-new-bot-law-strengthen-democracy

Collectif, Cerna. "Research Ethics in Machine Learning," PhD diss., CERNA; ALLISTENE, 2018.

Conitzer et al., “Moral Decision Making Frameworks for Artificial Intelligence”, 4834.

Coughlin, Joseph F., “The 'Internet of Things' Will Take Nudge Theory Too Far,” *Big Think*, March 27, 2017. https://bigthink.com/disruptive-demographics/the-internet-of-things-big-data-when-a-nudge-becomes-a-noodge.

Council of Europe, *European Social Charter (Revised)*, 3 May 1996, ETS 163. https://rm.coe.int/168007cf93.

Cowie, Roddie, “Ethical issues in affective computing,” In *The Oxford handbook of affective computing*, Oxford Library of Psychology, 2015.

Coyle, Stephen, Majidi, Carmel, LeDuc, Philip & Hsia, Jimmy, “Bio-Inspired Soft Robotics: Material Selection, Actuation, and Design”, *Extreme Mechanics* 22, July 2018.

Crawford, K., ‘A.I.'s White Guy Problem’, *The New York Times,* p. SR11, June 26, 2016. https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html.

Crawford, Matthew B., “Algorithmic Governance and Political Legitimacy,” *American Affairs*, Vol. III, No. 2, 2019.

Crouch, Will, “The most important unsolved problems in ethics,” 2012. http://blog.practicalethics.ox.ac.uk/2012/10/the-most-important-unsolvedproblems-in-ethics-or-how-to-be-a-high-impact-philosopher-part-iii/

Çürüklü, Baran, Dodig-Crnkovic, Gordana, & Akan, Batu, “Towards Industrial Robots with Human-like Moral Responsibilities”, *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference,* April 2010.

Cushman, Fiery, Liane Young, and Joshua Greene, “Multi-system moral psychology,” In *The moral psychology handbook*, J. M. Doris & the Moral Psychology Research Group, Eds., New York, NY: Oxford University Press, 2012, pp. 47–71.

Custers, Bart. "Drones Here, There and Everywhere Introduction and Overview." In *The Future of Drone Use*, pp. 3-20. TMC Asser Press, The Hague, 2016.

Daerden, Frank, and Dirk Lefeber. "Pneumatic artificial muscles: actuators for robotics and automation." *European journal of mechanical and environmental engineering* 47, no. 1 (2002): 11-21.

Dalkir, Knowledge Management in Theory and Practice, 217–244.

Dan Gettinger and Arthur Holland Michel, *Law Enforcement Robots Datasheet*, Center for the Study of the Drone, Bard College, 2016; Peter Asaro, ‘“Hands Up, Don’t Shoot!” HRI and the Automation of Police Use of Force’, *Journal of Human-Robot Interaction*, Vol. 5, No. 3, 2016, pp. 55–56.

Danaher, John, “Robotic Rape and Robotic Child Sexual Abuse: Should They be Criminalised?”, *Criminal Law and Philosophy*, December 2014.

Danks, David & London, Alex, “Regulating Autonomous Systems: Beyond Standards”, *IEEE Intelligent Systems*, 2017.

Dastin, Jeffrey, “Amazon scraps secret AI recruiting tool that showed bias against women,” *Reuters*, 10 October 2018. Retrieves at https://www.reuters.com/article/us-amazon-com-jobs-automation-in...-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

Datta, A., Tschantz M. C. & Datta, A. (2015). Automated Experiments on Ad Privacy Settings A Tale of Opacity, Choice, and Discrimination. *Proceedings on Privacy Enhancing Technologies*, 2015 (1), 92–112.

Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings. *Proceedings on privacy enhancing technologies*, *2015*(1), 92-112.

Daugherty, Paul R., and H. James Wilson, "How Humans and AI Are Working Together in 1,500 Companies," *Harvard Business Review*, April 4, 2019. https://hbr.org/2018/07/collaborative-intelligence-humans-and-ai-are-joining-forces.

David A. Graham, "The Dallas Shooting and the Advent of Killer Police Robots", *The Atlantic*, 8 July 2016. https://www.theatlantic.com/news/archive/2016/07/dallas-police-robot/490478/

David Grémillet et al., "Robots in Ecology: Welcome to the Machine," *Open Journal of Ecology*, Vol. 2, No. 2, 2012, p. 54

DeCew, Judith, "Privacy," In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Spring 2018 Edition)*, 2018. https://plato.stanford.edu/archives/spr2018/entries/privacy

Decker, Michael, Fischer, Martin & Ott, Ingrid, "Service Robotics and Human Labor: A First Technology Assessment of Substitution and Cooperation", *Elselvier Robotics and Autonomous Systems* 87, January 2017.

Dempsey, Caitlin, "Drones and GIS: A Look at the Legal and Ethical Issues", *GIS Lounge*, September 2015.

Denicolai, Lorzenzo, Grimaldi & Palmieri, Silvia, "Videos, Educational Robotics and Puppets: An Experimental Integration of Languages", Universita Degli Studi Di Torino, 2017.

Dennis, Louise, Michael Fisher, Marija Slavkovik, and Matt Webster, "Formal Verification of Ethical Choices in Autonomous Systems," *Robotics and Autonomous Systems,* Vol. 77, 2016, pp. 1–14., p. 1

Dhir, Amandeep, Yossatorn, Yossiri, Kaur, Puneet, & Chen, Sufen, "Online Social Media Fatigue and Psychological Wellbeing- A Study of Comulsive Use, Fear of Mission Out, Fatigue, Anxiety and Depression", *Elselvier International Journal of Information Management*, June 2018.

Diakopoulos, Nicholas, "Accountability in algorithmic decision making," *Communications of the ACM,* Vol. 59, No. 2, 2016, pp. 56-62.

Diakopoulos, Nicholas, "Accountability in Algorithmic Decision-Making," *Queue*, Vol. 13, No. 9, 2015, pp. 126–149.

Dillard and Yuthas, "Ethics Research in AIS".

Dobbe, R., Dean, S., Gilbert, T., and Kohli, N., 'A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics', 2018 Workshop on Fairness, Accountability and Transparency in Machine Learning during ICML 2018, Stockholm, Sweden, 2018. https://arxiv.org/abs/1807.00553.

Donovan, Joan, Robyn Caplan, Jeanna Matthews, "Algorithmic accountability: A primer. Data & Society Tech Algorithm Briefing: How Algorithms Perpetuate Racial Bias and Inequality," Washington, DC, USA, 2018. https://datasociety.net/wp-content/uploads/2018/04/Data_Society_Algorithmic_Accountability_Primer_FINAL.pdf

Doran, Derek, Sarah Schulz, and Tarek R. Besold, "What does explainable AI really mean? A new conceptualization of perspectives," *arXiv preprint arXiv:1710.00794*, 2017.

Dossett, Julian, "Artificial Intelligence: Raising New Ethics Questions in Media and Journalism," *PR Newswire for Journalists*, May 9, 2018. https://mediablog.prnewswire.com/2018/05/09/artificial-intelligence-ethics-questions/

Dressel, Julia and Farid, Hany, "The Accuracy, Fairness, and Limits of Predicting Recidivism," *Science Advances*, Vol. 4, 2018, p. 1.

Driessen, Clemens & Heutinck, Leonie F.M., "Cows Desiring to be Milked? Milking Robots and the Co-evolution of Ethics and Technology on Dutch Dairy Farms", *Agriculture and Human Values 32*(1), March 2015.,,,,,,

Droneseed, "Precision Forestry", accessed December 2018. droneseed.co

Duffy, Brian, "Fundamental Issues in Affective Intelligent Social Machines," *The Open Artificial Intelligence Journal*, No. 2, 2004, pp. 21–34.

Dumouchel, Paul & Damiano, Luisa, *Living with Robots*, 2017.

Dwork, Cynthia, and Aaron Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science,* Vol. 9, No. 3–4, 2014, pp. 211-407.

Dworkin, Gerald, *The theory and practice of autonomy*, Cambridge University Press, New York, 1988., 61f; Arneson, Richard, "Autonomy and Preference Formation," 1991, in Coleman, Jules L. and Allen Buchanan, eds, *In Harm's Way: Essays in Honor of Joel Feinberg*, Cambridge University Press, Cambridge, 1994, pp. 42–73.

Easton, Catherine, "Carry on Automat(r)on: Legal and Ethical Issues Relating to Healthcare Robots", *Tech Law*, May 2013.

Edlich, Alex & Sohoni, Vik, "Burned by the Bots: Why Robotic Autonomation is Stumbling", *McKinsey & Company*, May 2017.

EGE, *Future of Work, Future of Society,* 19 December 2018. https://ec.europa.eu/info/sites/info/files/research_and_innovation/ege/ege_future-of-work_opinion_122018.pdf

Elder, A., "Robot Friends for Autistic Children. Monopoly Money or Counterfeit Currency?," In Lin, P., Abney, K. and Jenkins, R. (eds.), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence.* Oxford University Press, 2017.

Elizabeth E. Joh, "Police Robots Need to Be Regulated to Avoid Potential Risks", *The New York Times*, 16 November 2016. https://www.nytimes.com/roomfordebate/2016/07/14/what-ethics-should-guide-the-use-of-robots-in-policing/police-robots-need-to-be-regulated-to-avoid-potential-risks

Elizabeth E. Joh, "Policing Police Robots", *UCLA Law Review Discourse*, Vol. 64, 2016, p. 521.

Elmaghraby & Losavio, 2014, p. 493

Elmaghraby, Adel S., and Michael M. Losavio, "Cyber security challenges in Smart Cities: Safety, security and privacy," *Journal of advanced research,* Vol. 5, No. 4, 2014, pp. 491-497., p. 492

Englisch, Joachim, "Digitalisation and the Future of National Tax Systems: Taxing Robots?," *Available at SSRN:* https://ssrn.com/abstract=3244670, September 5, 2018.

Eriksson, Alexander, and Neville Stanton, "The chatty co-driver: A linguistics approach to human-automation-interaction," In: *Contemporary ergonomics and human factors 2016: Proceedings of the international conference on ergonomics & human factors*, 2016.

Etzioni, Amitai & Etzioni, Oren, "Incorporating Ethics into Artificial Intelligence", *The Journal of Ethics* 21(4), December 2017.

European Commission for the Efficiency of Justice (CEPEJ), "European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment", adopted on 3-4 December 2018.

European Parliament, Public Undertakings and Services in the European Union. http://www.europarl.europa.eu/workingpapers/econ/w21/sum-2_en.htm

European Union, *Charter of Fundamental Rights of the European Union*, December 18, 2000, 2000/C 364/01. https://www.europarl.europa.eu/charter/pdf/text_en.pdf., art. 15.1

Eurostat, "Girls and women under-represented in ICT," 25 April 2018. Retrieved at https://ec.europa.eu/eurostat/web/products-eurostat-news/-/EDN-20180425-1.

Evans, Chadrick R., et al., "Telemedicine and telerobotics: from science fiction to reality", *Updates in Surgery*, Vol. 70, 2018, p. 361.

Fackler, Martin, "Six Years After Fukushima, Robots Finally Find Reactors' Melted Uranium Fuel", *New York Times*, November 2017.

Faggella, Daniel, "Machine Learning in Finanace—Present and Future Applications," *TechEmergence*, March 27, 2018. http://techemergence.com/machine-learning-in-finance-applications/

Fagnant, Daniel J. & Kockelman, Kara, "Preparing a Nation for Autonomous Vehicles: Opportunities, Barriers and Policy Recommendations," 2015.

Faniyi et al., "Architecting Self-Aware Software Systems".

Feiner, Lauren, "A woman shared her tragic story of how social media kept targeting her with baby ads after she had a stillbirth," *CNBC*, December 12, 2018, https://www.cnbc.com/2018/12/12/woman-calls-out-tech-companies-for-serving-baby-ads-after-stillbirth.html

Ferguson, Andrew Guthrie, “Predictive Prosecution,” *Wake Forest Law Review*, Vol. 51, 2016, pp. 705–44.

Ferguson, Andrew, *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*, New York University Press, New York, 2017.

Fessler, Leah, “We Tested Bots Like Siri and Alexa to See Who Would Stand Up to Sexual Harassment”, *Quartz*, February 2017.

Finn, Rachel & Donovan, Anna, “Big Data, Drone Data: Privacy and Ethical Impacts of the Intersection Between Big Data and Civil Drone Deployments”, *The Future of Drone Use*, October 2016.

Finn, Rachel & Wright, David, “Privacy, Data Protection and Ethics for Civil Drone Practice: A Survey of Industry, Regulators and Civil Society Organisations”, *Computer Law & Security Review 32*, 2016.

Finn, Rachel, David Wright, and Michael Friedewald, “Seven Types of Privacy,” In S. Gutwirth et al. (Eds.), *European Data Protection: Coming of Age*, Dordrecht: Springer, 2013.

Firth-Butterfield, K., Generation AI: What happens when your child's friend is an AI toy that talks back? *World Economic Forum* website, 22 May 2018. Retrieved at https://www.weforum.org/agenda/2018/05/generation-ai-what-happens-when-your-childs-invisible-friend-is-an-ai-toy-that-talks-back/.

Fiske, A., P. Henningsen, and A. Buyx, “Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy (Preprint)” doi: 10.2196/preprints.13216, 2018.

Fleetwood, Janet, “Public Health, Ethics, and Autonomous Vehicles,” *American Public Health Association*, 2017.

Fleming, Aysha, et al., “Is Big Data for Big Farming or for Everyone? Perceptions in the Australian Grains Industry,” *Agronomy for Sustainable Development*, Vol. 38, No. 24, 2018; Wolfert, Sjaak, et al., “Big Data in Smart Farming - A Review,” *Agricultural Systems*, Vol. 153, 2017.

Fletcher, S.R. & Webb, P., “Industrial Robot Ethics: Facing the Challenges of Human-Robot Collaboration in Future Manufacturing Systems”, *A World with Robots: International Conference on Robot Ethics 2015*, 2017.

Floyd, “The Paradigms of Programming”, 455.

Focused Ultrasound Therapy Using Robotic Approaches (FUTURA) Project, “Project Objectives”, accessed December 2018.

Fodor, *The Language of Thought*.

For more information see Barocas, Solon, and Andrew D. Selbst, “Big Datas Disparate Impact,” *Calif. L. Rev.,* Vol. 104, 2016, p. 671.

Foster, Malcolm, “Aging Japan: Robots May Have Role in Future of Elder Care”, *Reuters,* March 2018.

Foundation for Responsible Robotics, “Our Sexual Future with Robots”, 2017.

Fox, Maria, Derek Long, and Daniele Magazzeni, "Explainable planning," *arXiv preprint arXiv:1709.10256*, 2017.

Francis, Sam, “Robot Ethics: Three Things Industry Can Learn from New Robotic Standards”, *Robotics & Automation News*, March 2017.

Freeman, Tami, “Magnetic microbots line up for stem cell therapy”, *Physics World*, 30 May 2019. https://physicsworld.com/a/magnetic-microrobots-line-up-for-stem-cell-therapy/

Frey, Carl Benedikt, and Michael A. Osborne, "The future of employment: How susceptible are jobs to computerisation?," *Technological forecasting and social change,* Vol. 114, 2017, pp. 254-280. DOI: 10.1016/j.techfore.2016.08.019.

Frey, Carl Benedikt, and Osborne, Michael A., “The Future of Employment: How Susceptible Are Jobs to Computerisation?,” unpublished, September 2013, p. 3.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, *14*(3), 330-347.

Friedman, B., 1990. "Moral Responsibility and Computer Technology," Paper Presented at the Annual Meeting of the American Educational Research Association, Boston, Massachusetts.

Fuller, Daniel, Martine Shareck, and Kevin Stanley. "Ethical implications of location and accelerometer measurement in health research studies with mobile sensing devices." *Social Science & Medicine* 191 (2017): 84-88.

GAFA are Google, Apple, Facebook, and Amazon.

Gallagher, J., et al., "Junk News and Bots during the 2017 UK General Election: What Are UK Voters Sharing Over Twitter?," Data Memo, COMPROP-OII, May 2017. https://comprop.oii.ox.ac.uk/research/working-papers/junk-news-and-bots-during-the-2017-uk-general-election/

Gallego, Jelor, "The Microbots Will Treat Diseases From Inside Your Body", 9 Oct 2016. https://futurism.com/meet-the-microbots-that-will-treat-diseases-from-inside-your-body

Gams, Matjaz, Irene Yu-Hua Gu, Aki Härmä, Andrés Muñoz, and Vincent Tam, "Artificial Intelligence and Ambient Intelligence," *Journal of Ambient Intelligence and Smart Environments,* Vol. 11, No. 1, 2019, pp. 71-86., p. 76.

Garg, Vikas K., and Adam Tauman Kalai, "Meta-Unsupervised-Learning: A supervised approach to unsupervised learning," *arXiv preprint arXiv:1612.09030*, 2016.

Gavaghan, Colin, et al., "Government Use of Artificial Intelligence in New Zealand", New Zealand Law Foundation, Wellington, pp. 46–47, 2019.

Gennari et al., "The Evolution of Protégé".

Gevaert, Caroline, Richard Sliuzas, Claudio Persello, and George Vosselman. "Evaluating the societal impact of using drones to support urban upgrading projects." *ISPRS international journal of geo-information* 7, no. 3 (2018): 91.

Ghosh, Dipayan, "What is microtargeting and what is it doing in our politics?," *Internet Citizen*, October 4, 2018. https://blog.mozilla.org/internetcitizen/2018/10/04/microtargeting-dipayan-ghosh/

Giannelli, Gianna C., Lucia Mangiavacchi, and Luca Piccoli, "GDP and the value of family caretaking: how much does Europe care?," *Applied Economics,* Vol. 44, No. 16, 2012, pp. 2111-2131.

Giuliani, Manuel, Claus Lenz, Thomas Müller, Markus Rickert, and Alois Knoll. "Design principles for safety in human-robot interaction." *International Journal of Social Robotics* 2, no. 3 (2010): 253-274.

Glenn Greenwald, *No Place to Hide: Edward Snowden, the NSA and the Surveillance State*, Hamis Hamilton, London, 2014.

Goldstein, Jade, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell, "Summarizing text documents: sentence selection and evaluation metrics," In *SIGIR*, Vol. 99, no. 8, pp. 121-128, 1999.

Gonzalez-Fierro, Miguel, "10 Ethical Issues of Artificial Intelligence and Robotics", Github, April 2018.

Goodall, Noah J., "Ethical Decision Making During Automated Vehicle Crashes," *Transportation Research Record: Journal of the Transportation Research Board*, 2014.

Goodall, Noah, "From Trolleys to Risk: Models for Ethical Autonomous Driving", *American Journal of Public Healthy*, April 2017.

Goodman, Bryce, and Seth Flaxman, "European Union regulations on algorithmic decision-making and a "right to explanation"," *AI Magazine,* Vol. 38, No. 3, 2017, pp. 50-57.

Goodrich, Michael A., Dan R. Olsen, Jacob W. Crandall, and Thomas J. Palmer. "Experiments in adjustable autonomy." In *Proceedings of IJCAI Workshop on autonomy, delegation and control: interacting with intelligent agents*, pp. 1624-1629. Seattle, WA: American Association for Artificial Intelligence Press, 2001.

Gordon, John-Stewart, "What do we owe to intelligent Robots?," *AI & Society*, 2018.

Gorey, Colm, “Tiny robots in our blood could soon be used to sniff out and treat cancer”, *Silicon Republic*, 23 Nov 2017. https://www.siliconrepublic.com/machines/blood-cell-sized-robots-cancer

Graveleau, Séverin, “APB: Le gouvernement promet de se conformer aux demandes de la CNIL,” *Le Monde*, 28 September 2017. https://www.lemonde.fr/campus/article/2017/09/28/mise-en-demeure-de-la-cnil-pour-changer-le-fontionnement-d-admission-post-bac_5192758_4401467.html

Greenbaum, Dov, “Ethical, Legal and Social Concerns Relating to Exoskeletons”, *Computers and Society* 45(3), September 2015.

Greenbaum, Dov, “Ethical, Legal and Social Concerns Relating to Exoskeletons”, *Computers and Society* 45(3), September 2015.

Gronlun, Kirsten, “State of AI: Artificial Intelligence, the Military and Increasingly Autonomous Weapons”, *Future of Life Institute Website*, 9 May 2019. https://futureoflife.org/2019/05/09/state-of-ai/

Gunkel, David, “Mind the gap: responsible robotics and the problem of responsibility,” *Ethics of Information Technology*, 2017.

Gunkel, David, Robot Rights, The MIT Press, 2018.

Gunning, David, "Explainable artificial intelligence (xai)," *Defense Advanced Research Projects Agency (DARPA), nd Web,* Vol. 2, 2017.

Hall, Holly Kathleen, “Deepfake Videos: When Seeing Isn't Believing,” *Catholic University Journal of Law and Technology,* Vol. 27, No. 1, 2018, pp. 51-76. https://scholarship.law.edu/jlt/vol27/iss1/4.

Hamann, Heiko, Divband Soorati, Mohammad & Heinrich, Mary Katherine et al., “Flora Robotica— An Architectural System Combining Living Natural Plants and Distributed Robots”, *Cornell University Computer Science & Emerging Technologies*, September 2017.

Hanif, M. A., Khalid, F., Putra, R., Rehman, S., and Shafique, M., ‘Robust Machine Learning Systems: Reliability and Security for Deep Neural Networks’, 2018 IEEE 24th International Symposium on On-Line Testing And Robust System Design (IOLTS), 2-4 July 2018.

Hanson, F. A. (2009). Beyond the skin bag: On the moral responsibility of extended agencies. Ethics and Information Technology, 11(1), 91–99.

Harford, Tim, “Crash: How Computers are Setting us up for Disaster,” *The Guardian*, 2016. https://www.theguardian.com/technology/2016/oct/11/crash-howcomputers-are-setting-us-up-disaster

Harnden, Charlie, “How Artificial Intelligence is Destroying Meaningful Work,” *Medium*, https://medium.com/@charlieharnden/artificial-intelligence-and-meaningful-work-c8f6ec24f11b

Hart, Robert David, “If You’re Not a White Male, Artificial Intelligence’s Use in Healthcare Could Be Dangerous,” *QZ*, July 10, 2017. https://qz.com/1023448/if-youre-not-a-white-male-artificial-intelligences-use-in-healthcare-could-be-dangerous/

Hashimi, Ali, “AI Ethics: The Next Big Thing in Government. Anticipating the Impact of AI Ethics within the Public Sector”, World Government Summit; Deloitte, February 2019.

Havens, John, “Will we lose our rights as parents once robots are better at raising our kids?,” *Quarts*, July 10, 2019. Retrieved at https://qz.com/co/2533915/.

Hawksworth et al., 2018; World Economic Forum, *The Future of Jobs Report 2018*. Centre for the New Economy and Society, World Economic Forum, Switzerland, 2018a, http://www3.weforum.org/docs/WEF_Future_of_Jobs_2018.pdf.

Hawksworth, John, and Yuval Fertig, “AI and robots should create as many jobs as they displace in the long run,” *PricewaterhouseCoopers*, 2018. https://pwc.blogs.com/economics_in_business/2018/07/ai-and-robots-should-create-as-many-jobs-as-they-displace-in-the-long-run.html

Hawksworth, John, Euan Cameron, and Richard Berriman, “Will Robots Really Steal Our Jobs?: An International Analysis of the Potential Long-Term Impact of Automation,” *PricewaterhouseCoopers*, 2018. https://www.pwc.co.uk/economic-services/assets/international-impact-of-automation-feb-2018.pdf.

Hayes-Roth and Jacobstein, “The State of Knowledge-Based Systems”.

Helberger, Natali, Kari Karppinen, and Lucia D’acunto, “Exposure diversity as a design principle for recommender systems,” *Information, Communication & Society*, Vol. 21, No. 2, 2016, pp. 191-207. DOI: 10.1080/1369118X.2016.1271900.

Helbing, Dirk, et al., “Will Democracy Survive Big Data and Artificial Intelligence?,” *Scientific American*, 25 February 2017. https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/

Henderson, Peter, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau, "Ethical challenges in data-driven dialogue systems," In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 123-129, ACM, 2018.

High-Level Expert Group on Artificial Intelligence, “Ethics Guidelines for Trustworthy AI,” *European Commission*, July 4, 2019. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Hildebrandt, Mireille, “Chapter 14: Who is Profiling Who? Invisible Visibility.” In Gutwirth et al. (Eds.), *Reinventing Data Protection?*, Springer, 2009, pp.239-252.

Hildebrandt, Mireille, “The Meaning and The Mining of Legal Texts,” 2010. https://www.researchgate.net/publication/41463068_The_Meaning_and_the_Mining_of_Legal_Texts

Hildebrandt, Mireille, and Serge Gutwirth, "Concise conclusions: Citizens out of control," *Profiling the European Citizen*, Springer, Dordrecht, 2008, pp. 365-368.

Hildebrandt, Mireille, and Serge Gutwirth, *Profiling the European Citizen. Cross Disciplinary Perspectives*, Dordrecht: Springer, 2008.

Himmelreich, Johannes, “The Everyday Ethical Challenges of Self-Driving Cars”, *The Conversation*, March 2018.

Hodgson, Jarrod & Lian Pin Koh, “Best Practice for Minimising Unmanned Aerial Vehicle Disturbance to Wildlife in Biological Field Research”, *Current Biology* 26(10), May 2016.

Holstein, Tobias, Dodig-Crnkovic, Gordana & Pelliccione, Patrizio, “Ethical and Social Aspects of Self-Driving Cars,” *Cornell University*, 2018.

Hopkins, Anne, “The Ethical Debate on Drones”, *Digital Commons*, 2017.

Hornigold, Thomas, “Is the Rise of AI on Wall Street for Better or Worse?,” Singularity Hub, July 16, 2018. https://singularityhub.com/2018/07/16/is-the-rise-of-ai-on-wall-street-for-better-or-worse/

Hovy, Dirk, “Demographic Factors Improve Classification Performance,” *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015.

Hovy, Dirk, and Anders Søgaard, “Tagging Performance Correlates with Author Age,” *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2015.

Howard, Ayanna & Borenstein, Jason, “The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity”, *Science and Engineering Ethics* 24(5), October 2018.

Howard, Philip N., Woolley, Samuel, and Calo, Ryan, “Algorithms, Bots, and Political Communication in the US 2016 Election: The Challenge of Automated Political Communication for Election Law and Administration,” *Journal of Information Technology & Politics*, Vol. 15, No. 2, 2018, pp. 81–93

Howard, Philip, “How Political Campaigns Weaponize Social Media Bots ,” *IEEE Spectrum*, October 18, 2018. https://spectrum.ieee.org/computing/software/how-political-campaigns-weaponize-social-media-bots

Hu, Schroeder, and Starr, “A Knowledge-Based Real-Time Diagnostic System for PLC Controlled Manufacturing Systems”.

Hulme, David, “Rogue Robots”, *Vision Insight: Global Threats*, August 2018.

Hurlburt, George, “How Much Should We Trust Artificial Intelligence,” *InfoQ*, September 8, 2017. https://www.infoq.com/articles/ai-trust/

IBM, “With AI, our words will be a window into our mental health”, *IBM*, n.d. https://www.research.ibm.com/5-in-5/mental-health/

Ienca, Marcello, Tenzin Wangmo, Fabrice Jotterand, Reto W. Kressig, and Bernice Elger, "Ethical design of intelligent assistive technologies for dementia: a descriptive review," *Science and engineering ethics,* Vol 24, No. 4, 2018, pp. 1035-1055.

Irene Y. Chen, Szolovits, Peter, and Ghassemi, Marzyeh, “Can AI Help Reduce Disparities in General Medical and Mental Health Care,” *AMA Journal of Ethics*, Vol. 21, No. 2, 2019.

Ivošević, Bojana, Han, Yong-Gu, Cho, Youngho & Kwon, Ohseok, “Monitoring Butterflies With an Unmanned Aerial Vehicle: Current Possibilities and Future Potentials”, *Journal of Ecology and Environment 41*(1), 2017.

Ivošević, Bojana, Han, Yong-Gu, Cho, Youngho & Kwon, Ohseok, “The Use of Conservation Drones in Ecology and Wildlife Research”, *Ecology and Environment*, *38*(1), February 2015.

Jansen, Philip, Stearns Broadhead, Rowena Rodrigues, David Wright, Philip Brey, Alice Fox, Ning Wang, *SIENNA D4.1 State-of-the-art Review*, WP4 - AI & Robotics, 2018, Public deliverable report from the SIENNA project. http://www.sienna-project.eu/publications/deliverable-reports/

Jenna Burrell, (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1).

Jensen, Ole B., “Drone City—Power, Design, and Aerial Mobility in the Age of Smart Cities”, April 2016.

Job loss due to AI — How bad is it going to be? (2019, February 4). *Skynet Today.* Retrieved at: https://www.skynettoday.com/editorials/ai-automation-job-loss.

John Arquilla and David Ronfeldt, “Swarming & The Future of Conflict”, RAND, National Defense Research Institute, 2000. See also section on use of AI and robotics in the defence sector in the present report.

Johnson, J. A. (2018). Open Data, Big Data, and Just Data. In *Toward Information Justice*, ed. by J. A. Johnson, 23-49.

Johnson, Khari, “How AI Can Strengthen and Defend Democracy,” *Venture Beat*, 4 July 2019. https://venturebeat.com/2019/07/04/how-ai-can-strengthen-and-defend-democracy/

Joni R. Jackson, (2018). Algorithmic Bias. Journal of Leadership, Accountability and Ethics, 15(4), 55-65.

Jordan, M. I., and T. M. Mitchell, “Machine Learning: Trends, Perspectives, and Prospects,” *Science,* Vol. 349, No. 6245, 2015, pp. 255–260.

Jordan, M. I., and T. M. Mitchell, “Machine Learning: Trends, Perspectives, and Prospects,” *Science,* Vol. 349, No. 6245, 2015, pp. 255–260.

Kahn Jr. Peter H., Kanda, Takayuki, & Ishiguro, Hiroshi et al., “Do People Hold a Humanoid Robot Morally Accountable for the Harm it Causes?”, *Attitudes and Responses to Social Robots*, March 2012.

Kahn Jr., Peter H., Kanda, Takayuki, & Ishiguro, Hiroshi, et al., “’Robovie, You’ll Have to Go into the Closet Now’: Children’s Social and Moral Relationships with a Humanoid Robot”, *Developmental Psychology* 48(2), 2012.

Kamishima, Akaho, & Sakuma, 2011, p. 644; see also Dwork, Cynthia, and Aaron Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science,* Vol. 9, No. 3–4, 2014, pp. 211-407.

Kamishima, T., Akaho, S., & Sakuma, J. (2011, December). Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops* (pp. 643-650). IEEE.

Kaushik, Preetam, "Is Artifical Intelligence the way Forward for Personal Finance," *Wired.* http://wired.com/insights/2014/02/artificial-intelligence-way-forward-personal-finance/

Keeling, Geoff, "Legal Necessity, Pareto Efficiency & Justified Killing in Autonomous Vehicle Collisions", *Ethical Theory and Moral Practice* 21(2), April 2018.

Kemp, Luke, "In the Age of Deepfakes, Could Virtual Actors Put Humans out of Business?," *The Guardian*, Guardian News and Media, July 8, 2019. https://www.theguardian.com/film/2019/jul/03/in-the-age-of-deepfakes-could-virtual-actors-put-humans-out-of-business

Keyes, O., 'The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition', *Proceeding of the ACM on Human-Computer Interaction*, Vol. 2, No. CSCW, Article 88, November 2018.

King, T. C., Aggarwal, N., Taddeo, M., and Floridi, L., 'Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions', *Science and Engineering Ethics*, 2019. https://doi.org/10.1007/s11948-018-00081-0.

Kirschgens, Laura, Ugarte, Irati & Uriarte, Endika et al., "Robot Hazards: From Safety to Security", Whitepaper, 2018.

Kitano, Hiroaki, "Artificial intelligence to win the nobel prize and beyond: Creating the engine for scientific discovery," *AI magazine* Vol. 37, No. 1, 2016, pp. 39-49. DOI: https://doi.org/10.1609/aimag.v37i1.2642.

Kitchin, Rob, "The real-time city? Big data and smart urbanism," *GeoJournal,* Vol. 79, No. 1, 2014, pp. 1-14.

Kladitis, Paul E., "How small is too small? True microrobots and nanorobots for military applications in 2035", Research Report, Maxwell Air Force Base, Alabama, April 2010 and see also references in the section on Swarm robots.

Knight, Kevin, and Irene Langkilde, "Preserving ambiguities in generation via automata intersection," *AAAI/IAAI*, pp. 697-702, 2000.

Knuth. Donald. "Computer Science and Its Relation to Mathematics," *American Mathematical Monthly*, Vol. 81, No. 4, 1974, pp. 323-343.

Kobie, Nicole, "The questionable ethics of treating autistic children with robots," *WIRED*, July 18, 2018. https://www.wired.co.uk/article/autisim-children-treatment-robots

Koene, Ansgar, Chris Clifton, Yohko Hatada, Helena Webb, Menisha Patel, Caio Machado, Jack LaViolette, Rashida Richardson, and Dillon Reisman, "A governance framework for algorithmic accountability and transparency," *European Parliamentary Research Service*, 2019. http://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_STU(2019)624262

Kohonen, "Exploration of Very Large Databases by Self-Organizing Maps".

Korinek, Anton, "Labor in the Age of Automation and Artificial Intelligence," *Economics for Inclusive Prosperity*, February 2019. https://econfip.org/policy-brief/labor-in-the-age-of-automation-and-artificial-intelligence/.

Korinek, Anton, and Joseph E. Stiglitz, "Artificial intelligence and its implications for income distribution and unemployment," in Goldfarb, Avi, Joshua Gans, and Ajay Agrawal (eds.), *The Economics of Artificial Intelligence: An Agenda*, University of Chicago Press, 2019.

Körtner, T., "Ethical challenges in the use of social service robots for elderly people," *Zeitschrift Für Gerontologie Und Geriatrie*, Vol. 49, No. 4, 2016, pp. 303–307.

Krach, Sören, Frieder M. Paulus, Maren Bodden, and Tilo Kircher, "The Rewarding Nature of Social Interactions," *Frontiers in behavioral neuroscience,* Vol. 4, p. 22, 2010.

Kraemer, Felicitas, Kees van Overveld, and Martin Peterson, “Is there an ethics of algorithms?,” *Ethics and Information Technology*, Volume 13, Issue 3, 2011, pp. 251–260.

Kristian Lum and James Johndrow, ‘A Statistical Framework For Fair Predictive Algorithms’, 2016, 1 <https://arxiv.org/abs/1610.08077> [accessed 1 July 2019].

Kroeker, K. L., ‘Graphics and Security: Exploring Visual Biometrics’, *IEEE Computer Graphics and Applications*, Vol. 22, No. 4, July 2002, pp. 16-21.

Kulić, Dana, and Elizabeth Croft. "Pre-collision safety strategies for human-robot interaction." *Autonomous Robots*22, no. 2 (2007): 149-164.

Kulkarni, Anagha, Chakraborti, Tathagata & Zha, Yantian et al., “Explicable Robot Planning as Minimized Distance from Expected Behavior”, *Cornell University,* July 2018.

Lachow, Irving, “The Upside and Downside of Swarming Drones,” *Bulletin of the Atomic Scientists*, Vol. 73, No. 2, 2017, p. 96; Bredeche, Nicolas, Haasdijk, Evert, and Prieto, Abraham, “Embodied Evolution in Collective Robotics: A Review,” *Frontiers in Robotics and AI*, Vol. 5, No. 12, 2018; Magnuson, Stew, “Military Beefs Up Research Into Swarming Drones,” *National Defense Magazine*, March 1, 2016. https://www.nationaldefensemagazine.org/articles/2016/2/29/2016march-military-beefs-up-research-into-swarming-drones

Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind.* University of Chicago Press.

Lambrecht, Anja and Catherine Tucker, “Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads,” *SSRN*, 2018, retrievable at https://ssrn.com/abstract=2852260 or http://dx.doi.org/10.2139/ssrn.2852260.

Land, Amjad, and Nolas, “Accountability and Ethics in Knowledge Management”, 2.

Langheinrich, Marc, “Privacy by Design — Principles of Privacy-Aware Ubiquitous Systems,” *Ubicomp 2001: Ubiquitous Computing Lecture Notes in Computer Science*, 2001, pp. 273–291., p. 6.

Lau, Josephine, Benjamin Zimmerman, and Florian Schaub, "Alexa, are you listening?: Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers," *Proceedings of the ACM on Human-Computer Interaction,* Vol 2, No. CSCW, 2018, p. 102.

Leavy, Susan, “Gender bias in artificial intelligence: the need for diversity and gender theory in machine learning,” *2018 ACM/IEEE 1st International Workshop on Gender Equality in Software Engineering*. DOI: 10.1145/3195570.3195580.

Lee, Byounggwan, Ohkyun Kwon, Inseong Lee, and Jinwoo Kim, "Companionship with smart home devices: The impact of social connectedness and interaction types on perceived social support and companionship in smart homes," *Computers in Human Behavior,* Vol 75, 2017, pp. 922-934.

Lee, J. D., and See, K. A., ‘Trust in Automation: Designing for Appropriate Reliance’, *Human Factors*, Vol. 46, No. 1, pp. 50-80, Spring 2004.

Leggett, Theo, “Who is to Blame for ‘Self-Driving Car’ Deaths?”, BBC, May 2018.

Lei, Tao, Regina Barzilay, and Tommi Jaakkola, “Rationalizing Neural Predictions,” *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.

Leidner, Jochen L., and Vassilis Plachouras, “Ethical by Design: Ethics Best Practices for Natural Language Processing,” *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 2017.

Lenca, Marcello, Kressig, Reto & Jotterand, Fabrice et al., “Proactive Ethical Design for Neuroengineering, Assistive and Rehabilitation Technologies: The Cybathlon Lesson”, Journal of Neuroengineering *Rehabilitation* 14, November 2017.

Li, Jamy, “The benefit of being physically present: A survey of experimental works comparing co-present robots, telepresent robots and virtual agents,” *International Journal of Human-Computer Studies*, Vol. 77, 2015, pp. 23–37.

Lidynia, Chantal, Philipsen, Ralf & Ziefle, Martina, “Droning on About Drones—Acceptance of and Perceived Barriers to Drones in Civil Usage Contexts”, *Advances in Human Factors in Robots and Unmanned Systems*, 2017.

Lin, Lucas, and Shmueli, “Research Commentary —Too Big to Fail”, 9–10.

Lin, Patrick, “Drone-Ethics Briefings: What a Leading Robot Expert Told the CIA,” *The Atlantic*, December 21, 2011.

Lin, Patrick, “Why Ethics Matters for Autonomous Cars”, *Autonomous Driving: Technical, Legal and Social Aspects*, 2016.

Lin, Patrick, Abney, Keith, and Bekey, George, “Robot Ethics: Mapping the Issues for a Mechanized World”, *Artificial Intelligence*, Vol. 175, 2011, p. 947.

Lin, Patrick, Keith Abney and George A. Bekey, “Current Trends in Robotics: Technology and Ethics,” Robot Ethics: The Ethical and Social Implications of Robotics, MIT Press, 2012.

Lindsey Barrett, “Reasonably Suspicious Algorithms: Predictive Policing at the United States Border”, *N.Y.U. Review of Law and Social Change*, Vol. 41, 2017, p. 343.

Lipton, Zachary C., “The Mythos of Model Interpretability,” *Communications of the ACM*Vol. 61, No. 10, 2018, pp. 36–43.

Lipton, Zachary C., “The Mythos of Model Interpretability,” *Communications of the ACM,* Vol. 61, No. 10, 2018, pp. 36–43.; Doshi-Velez, Finale, and Been Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

Liu, Hui, Yinghui Huang, Zichao Wang, Kai Liu, Xiangen Hu, and Weijun Wang, “Personality or Value: A Comparative Study of Psychographic Segmentation Based on an Online Review Enhanced Recommender System,” *Applied Sciences*, Vol. 9, No. 10, 2019, pp. 1992, DOI:10.3390/app9101992.

Lubin, Gus, “The Incredible Story Of How Target Exposed A Teen Girl's Pregnancy,” *Business Insider*, February 12, 2012, https://www.businessinsider.com/the-incredible-story-of-how-target-exposed-a-teen-girls-pregnancy-2012-2?international=true&r=US&IR=T.

Lutz, Christoph & Tamò, Aurelia, “Privacy and Healthcare Robots—An ANT Analysis”, 2016.

Lyria Bennett Moses and Janet Chan, ‘Algorithmic Prediction in Policing: Assumptions, Evaluation, and Accountability’, *Policing and Society*, Vol. 28, No. 7, 2018, p. 806.

Macnish, K., ‘Unblinking Eyes: The Ethics of Automating Surveillance’, *Ethics and Information Technology*, Vol. 14, No. 2, pp. 151-167, June 2012.

Mainbot, “Industrial Objectives”, accessed December 2018.

Manjikian, Mary, “A Typology of Arguments About Drone Ethics”, *Strategic Studies Institute US Army War College*, October 2017.

Manon Oostveen and Irion, K. (2018). ‘The Golden Age of Personal Data: How to Regulate an Enabling Fundamental Right?’ in M. Bakhoum, B. Conde Gallego, M-O. Mackenrodt, & G. Surblytė-Namavičienė (Eds.), *Personal Data in Competition, Consumer Protection and Intellectual Property Law: Towards a Holistic Approach?* (pp. 7-26). (MPI Studies on Intellectual Property and Competition Law; Vol. 28). Berlin: Springer.

Manuel Souto-Otero and Benito-Montagut, R. (2016). ‘From governing through data to governmentality through data: Artefacts, strategies and the digital turn’ in *European Educational Research Journal* Vol. 15(1): 14-33; Luke Hutton & Henderson, T. (2017). ‘Beyond the EULA: Improving Consent for Data Mining’ in *Transparent Data Mining for Big and Small Data*, (Ed.) Tania Cerquitelli, Daniele Quercia and Frank Pasquale. Springer.

Manyika, James, Susan Lund, Michael Chui, Jacques Bughin, Jonathan Woetzel, Parul Batra, Ryan Ko, and Saurabh Sanghvi, “Jobs Lost, Jobs Gained: Workforce Transitions in a Time of Automation,” *McKinsey Global Institute*, 2017.

Mariano, Pedro, Salem, Ziad & Mills, Rob et al., “Design Choice for Adapting Bio-Hybrid Systems with Evolutionary Computation”, *GECCO 2017 Companion*, July 2017.

Mark Coeckelbergh, “From Killer Machines to Doctrines and Swarms, or Why Ethics of Military Robotics Is Not (Necessarily) About Robots,” *Philosophy & Technology*, Vol. 24, 2011, p. 273.

Markou, Christopher, “Why Using AI to Sentence Criminals Is a Dangerous Idea,” *The Conversation*, May 2017. http://theconversation.com/why-using-ai-to-sentence-criminals-is-a-dangerous-idea-77734

Martin, Ron, and Philip S. Morrison (eds.), *Geographies of labour market inequality*, Routledge, London and New York, 2003.

Marx, Paris, “Humans to serve the rich, robots to serve the poor,” *Medium*, August 28, 2016. https://medium.com/radical-urbanist/humans-to-serve-the-rich-robots-to-serve-the-poor-6e2efc95c1b4

Mason, “Four Ethical Issues of the Information Age”.

Matthias, Andreas, “The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata,” *Ethics and Information Technology,* Vol. 6, No. 3, 2004, pp. 175–183.

Matz, Sandra C., Michal Kosinski, Gideon Nave, and David J. Stillwell, “Psychological targeting as an effective approach to digital mass persuasion,” *Proceedings of the National Academy of Sciences of the United States of America* Vol. 114, No. 48, 2017, pp. 12714–12719.

Maurice, Pauline, Allienne, Ludivine & Malaise, Adrien et al, “Ethical and Social Considerations for the Introduction of Human-Centered Technologies at Work”, *IEEE Workshop on Advanced Robotics and its Social Impacts*, 2018.

Mavroforou, A., Michalodimitrakis, E., Hatzitheo-Filou, C., & Giannoukas, A. “Legal and Ethical Issues in Robotic Surgery”, *PubMed*, February 2010.

Maxmen, Amy, “Self-Driving Car Dilemmas Reveal that Moral Choices are Not Universal”, *Nature*, October 2018.

McFarland, Matt, “DARPA’s Robotics Challenge has a gender problem.” *Washington Post*, June 5, 2015. Retrieved at http://www.washingtonpost.com/blogs/innovations/wp/2015/06/05/darpas-robotics-challenge-has-a-gender-problem/.

Meghdari, Ali & Alemi, Minoo, “Recent Advances in Social & Cognitive Robotics and Imminent Ethical Challenges”, *Advances in Social Science, Education, and Humanities Research* 211, 2018.

Mehr, Hila, “Artificial Intelligence for Citizen Services and Government”, Harvard Ash Center for Democratic Governance and Innovation, August 2017.

Meinecke, Lisa & Voss, Laura, “I Robot, You Unemployed: Robotics in Science Fiction and Media Discourse”, *Schafft Wissen: Gemeinsames und Geteiltes Wissen in Wissenschaft und Technik*, pp.203-215, October 2016.

Microsoft, “Healthcare, Artificial Intelligence, Data and Ethics – A 2030 Vision How responsible innovation can lead to a healthier society”, December 2018. https://www.digitaleurope.org/wp/wp-content/uploads/2019/02/Healthcare-AI-Data-Ethics-2030-vision.pdf

Mieskes, Margot, “A Quantitative Study of Data in the NLP Community,” *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 2017.

Milano, Silvia, Mariarosaria Taddeo, and Luciano Floridi, “Recommender Systems and their Ethical Challenges,” 2019. http://dx.doi.org/10.2139/ssrn.3378581

Miller, Christopher A., Harry Funk, Robert Goldman, John Meisner, and Peggy Wu, "Implications of adaptive vs. adaptable UIs on decision making: Why “automated adaptiveness” is not always the right answer," In *Proceedings of the 1st international conference on augmented cognition*, pp. 22-27. 2005., p. 3

Miller, T., ‘Explanation in Artificial Intelligence: Insights From the Social Sciences’, *Artificial Intelligence*, Vol. 267, pp. 1-38, February 2019.

Miller, Tim, Piers Howe, and Liz Sonenberg, "Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences," *arXiv preprint arXiv:1712.00547*, 2017.

Mitchell, Anna and Diamond, Larry, "China's Surveillance State Should Scare Everyone," *The Atlantic*, 2 February 2018. https://www.theatlantic.com/international/archive/2018/02/china-surveillance/552203/

Mitchell, T. M. (2006). *The discipline of machine learning* (Vol. 9). Pittsburgh, PA: Carnegie Mellon University, School of Computer Science, Machine Learning Department.

Mitchell, Tom M., *Machine Learning*, McGraw Hill, New York, 1997.

Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S. and Floridi, L. (2016). The Ethics of Algorithms: Mapping the Debate.

Mladenovic, Milos N. & McPherson, Tristram, "Engineering Social Justice into Traffic Control for Self-Driving Vehicles?" *Science and Engineering Ethics*, August 2016.

Morgan, Blake, "How Artificial Intelligence Will Impact the Insurance Industry," *Forbes,* July 25, 2017. http://forbes.com/sites/blakemorgan/2017/07/25/how-artificial-intelligence-will-impact-the-insurance-industry/#5255ab2e6531

Mortensen, Dennis, "Automation May Take Our Jobs—But It'll Restore Our Humanity", *Quartz Automation Revolution*, August 2017.

Mouthuy, Pierre-Alexis & Carr, Andrew, "Growing Tissue Grafts on Humanoid Robots: A Future Strategy in Regenerative Medicine?" *Science Robotics* 2(4)*,* March 2017.

Moyle, Wendy, Bramble, Marguerite, Jones, Cindy & Murfield, Jenny, "'She Had a Smile on Her Face as Wide as the Great Australian Bite': A Qualitative Examination of Family Perceptions of a Therapeutic Robot and a Plush Toy", *The Gerontologist* 00(00), October 2017.

Moyle, Wendy, Bramble, Marguerite, Jones, Cindy & Murfield, Jenny, "Care Staff Perceptions of a Social Robot Called Paro and a Look-Alike Plush Toy: A Descriptive Qualitative Approach", *Aging & Mental Health* 22(3), November 2016.

Müller, V. C. (ed.), *Risks of Artificial Intelligence*, Boca Raton, CRC Press, 2016; Bostrom, N., *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, 2014.

Müller, Vincent C., and Nick Bostrom, "Future Progress in Artificial Intelligence: A Survey of Expert Opinion," Fundamental Issues of Artificial Intelligence, 2016, pp. 555–572.

Mulligan, Christina, "Revenge Against Robots", Brooklyn Law School, 2017.

Murashov, Vladamir, Hearl, Frank & Howard, John, "Working Safely With Robot Workers: Recommendations for the New Workplace", *Journal of Occupational and Environmental Hygiene*, *13*(3), 2016.

Murison, Malek, "The Great Firewall: China looks to AI to censor online material," *Internet of Business*, May 23, 2018. https://internetofbusiness.com/china-censorship-online-material-ai/

Muro, Mark, Robert Maxim, and Jacob Whiton, "Automation and Artificial Intelligence: How Machines are Affecting People and Places," *Brookings Institution*, https://www.brookings.edu/research/automation-and-artificial-intelligence-how-machines-affect-people-and-places/, 2019.

Mussolum, Erin, "How Art Shapes Identity", Trinity Western University*,* October 2007.

Myers, Andrew, "An artificial intelligence algorithm developed by Stanford researchers can determine a neighborhood's political leanings by its cars," *Stanford News*, November 28, 2017. https://news.stanford.edu/2017/11/28/neighborhoods-cars-indicate-political-leanings/

Nan, Khan, and Iqbal, "Real-Time Fault Diagnosis Using Knowledge-Based Expert System".

Naser and Zaqout, "Knowledge-Based Systems That Determine the Appropriate Students Major".

National Science Board, *Science & Engineering Indicators 2018*. Retrieved at https://nsf.gov/statistics/2018/nsb20181/report.

Nau, Dana S, "Current trends in automated planning," *AI magazine,* Vol. 28, no. 4 (2007): 43-58., p. 43

Nedelkoska & Quintini, 2018

Nedelkoska, Ljubica, and Glenda Quintini, "Automation, skills use and training," *OECD Social, Employment and Migration Working Papers*, No. 202, OECD Publishing, Paris, 2018, http://dx.doi.org/10.1787/2e2f4eea-en.

NEM, “Artificial Intelligence in the Media and Creative Industries,” Position paper, July 2018, https://nem-initiative.org/wp-content/uploads/2018/10/nem-positionpaper-aiinceativeindustry.pdf

Neudert, Lisa Maria, and Marchal, Nahema, “Polarisation and the Use of Technology in Political Campaigns and Communication,” Study Panel for the Future of Science and Technology, European Parliamentary Research Service, Brussels, March 2019.,,,http://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_STU(2019)634414

Nezhadali, Vaheed, “Multi-Objective Optimization of Industrial Robots”, Linköpings Universitet, 2011.

Nissenbaum, Helen, “How computer systems embody values,” *Computer*, Vol. 34, No. 3, 2001, pp. 120–119.

Nitzberg, Mark, Olaf Groth, and Mark Esposito, “AI Isn't Just Compromising Our Privacy-It Can Limit Our Choices, Too,” *Quartz*, December 13, 2017. https://qz.com/1153647/ai-isnt-just-taking-away-our-privacy-its-destroying-our-free-will-too/.

Norval, A., and Prasopoulou, E., ‘Public Faces? A Critical Exploration of the Diffusion of Face Recognition Technologies in Online Social Networks’, *New Media & Society*, Vol. 19, No. 4, pp. 637-654, April 2017.

Nunez, Catherine, “Artificial Intelligence and Legal Ethics: Whether AI Lawyers Can Make Ethical Decisions,” *Tulane Journal of Technology and Intellectual Property*, Vol. 20, 2017, pp. 189–204

O’Neil, *Weapons of Math Destruction*, London, Allen Lane, 2016.

Oda, Joanna. Fong, Daniel. Zitouni, Abderrachid and Kosatsky, Tom. “Health Effects of Large LED Screens on Local Residents.” National Collaborating Centre for Environmental Health. http://www.ncceh.ca/documents/practice-scenario/health-effects-large-led-screens-local-residents (retrieved 01/06/2019)

Odhran James McCarthy, “AI and Global Governance: Turning the Tide on Crime with Predictive Policing”, Center for Policy Research, United Nation University, February 2019 https://cpr.unu.edu/ai-global-governance-turning-the-tide-on-crime-with-predictive-policing.html [accessed 6 July 2019].

OSF Journalism, “Artificial intelligence demands genuine journalism,” *Medium*, October 31, 2018. https://medium.com/innovation-in-journalism/artificial-intelligence-demands-genuine-journalism-8519c4e0fc86

Oudshoorn, Nelly, E. Rommes, and E. Stienstra, “Configuring the User as Everybody: Gender and Design Cultures in Information and Communication Technologies,” *Science Technology Human Values*, Vol. 29, No. 1, 2004, pp. 30–63.

Owais Quereshi, Mohammed & Sajjad Syed, Rumaiya, “The Impact of Robotics on Employment and Motivation of Employees in the Service Sector, with Special Reference to Health Care”, *Oshri Saftey and Health at Work* 5(4), December 2014

P. Lichocki, P. Kahn Jr, and A. Billard. The Ethical Landscape of Robotics. *IEEE Robotics and Automation Magazine*, 18(1):39-50, 2011. (Cited as requested)

Pachidis, Theodore, Vrochidou, Eleni, & Kaburlasos, Vassilis et al., “Social Robotics in Education: Stat-of-the-Art and Directions, 27th International Conference on Robotics RAAD, July 2018

Pagallo, Ugo. "Robots in the cloud with privacy: A new threat to data protection?." *Computer Law & Security Review* 29, no. 5 (2013): 501-508.

Papernot, N., McDaniel, P., Sinha, A., and Wellman, M., 'Towards the Science of Security and Privacy in Machine Learning', p. 4, 2016. https://arxiv.org/abs/1611.03814.

Papernot, Nicolas, Patrick Mcdaniel, Arunesh Sinha, and Michael P. Wellman, "SoK: Security and Privacy in Machine Learning," *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, 2018.

Pariser, Eli, *The Filter Bubble: What the Internet Is Hiding from You*, Penguin Books, London, 2011.

Parsons, E. C. M., Sarah J. Dolman, Andrew J. Wright, Naomi A. Rose, and W. C. G. Burns. "Navy sonar and cetaceans: Just how much does the gun need to smoke before we act?." *Marine pollution bulletin* 56, no. 7 (2008): 1248-1257.

Passchier-Vermeer, Willy, and Wim F. Passchier. "Noise exposure and public health." *Environmental health perspectives* 108, no. suppl 1 (2000): 123-131.

Patrick Lin, Keith Abney and George Bekey, "Robot Ethics: Mapping the Issues for a Mechanized World", *Artificial Intelligence*, vol. 175, 2011, p. 947.

Patterson, Dan, "How AI-Powered Robots Will Protect the Networked Soldier," *TechRepublic*, April 6, 2016. https://www.techrepublic.com/article/how-ai-powered-robots-will-protect-the-networked-soldier.

Pearson, Yvette & Borenstein, Jason, "Creating 'Companions' for Children: The Ethics of Designing Esthetic Features for Robots", *AI & Society*, February 2014.

Peertechz Journal, Engineering Group, "Aims and Scope", *Annals of Robotics and Automation*, accessed December 2018.

PERVADE: Pervasive Data Ethics, ""The study has been approved by the IRB": Gayface AI, research hype and the pervasive data ethics gap," *Medium (PERVADE: Pervasive Data Ethics)*, November 30, 2018. https://medium.com/@pervade_team/the-study-has-been-approved-by-the-irb-gayface-ai-research-hype-and-the-pervasive-data-ethics-3b36c5a53eec

Petropoulos, Georgios, "The Impact of Artificial Intelligence on Employment", in Neufeind, Max, O'Reilly, Jacqueline, Ranft, Florian, *Work in the digital age*, Rowman and Littlefield, London, 2018, pp. 119-132.

Pfeifle, Anne, "Alexa, What Should We Do about Privacy: Protecting Privacy for Users of Voice-Activated Devices," *Wash. L. Rev,* Vol 93, 2018, pp. 421-458.

Pinar Saygin, Ayse, Chaminade, Thierry & Ishiguro, Hiroshi et al., "The Thing That Should Not Be: Predictive Coding and the Uncanny Valley in Perceiving Human and Humanoid Robot Actions", *Social Cognitive and Affective Neuroscience* 7(4), April 2012.

Pisch, Anita, "The Ethics of Human Robots: Sam Jinks Brings an Artist's Perspective to the Discourse", *The Conversation*, October 2017.

Polgar, David Ryan, "Is it Unethical to Design Robots to Resemble Humans?", *Quartz*, June 2017.

Popenici, Stefan A., and Sharon Kerr, "Exploring the impact of artificial intelligence on teaching and learning in higher education," *Research and Practice in Technology Enhanced Learning,* Vol. 12, No. 1, 2017, p. 22. DOI 10.1186/s41039-017-0062-8.; Johnson, Jeffrey Alan, "The ethics of big data in higher education," *International Review of Information Ethics,* Vol. 21, No. 21, 2014, pp. 3-10. http://www.i-r-i-e.net/inhalt/021/IRIE-021-Johnson.pdf

Powles and Hodson, **REMOVE**, 2017; Nuffield Council on Bioethics, "Artificial Intelligence (AI) in Healthcare and Research", May 2018, p. 2.

Powles, Julia, and Hodson, Hal, "Google DeepMind and Healthcare in an Age of Algorithms," *Health and Technology*, Vol. 7, 2017; Forbes Insights, "Rethinking Medical Ethics," February 2019. https://www.forbes.com/sites/insights-intelai/2019/02/11/rethinking-medical-ethics/

Prabhakar, S., Pankanti, S., and Jain, A. K., 'Biometric Recognition: Security and Privacy Concerns', *IEEE Security & Privacy*, Vol. 1, No. 2, pp. 33-42, March-April 2003.

Preece, Alun, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty, "Stakeholders in explainable AI," *arXiv preprint arXiv:1810.00184*, 2018s.

Prescott, Tony, Lepora, Nathan & Verschure, Paul (eds.), *Living Machines: A Handbook of Research in Biomimetic and Biohybrid Systems*, Oxford University Press, 2018.

Privacy International, "Artificial Intelligence", *Privacy International*, n.d. https://privacyinternational.org/topics/artificial-intelligence

Pryzant, Reid, Kelly Shen, Dan Jurafsky, and Stefan Wagner, "Deconfounded Lexicon Induction for Interpretable Social Science," *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018., p. 1.

Radford, A., Wu, J., Amodei, D., Amodei, D., Clark, J., Brundage, M., and Sutskever, I., 'Better Language Models and Their Implications', OpenAI, February 14, 2019. https://openai.com/blog/better-language-models/.

Raisinghani, Mahesh S., Ally Benoit, Jianchun Ding, Maria Gomez, Kanak Gupta, Victor Gusila, Daniel Power and Oliver Schmedding, "Ambient intelligence: Changing forms of human-computer interaction and their social implications," *Journal of digital information,* Vol. 5, No. 4, 2004.

Rajkomar, Alvin, "Scalable and accurate deep learning with electronic health records," *NJP Digital Magazine*, 2018. https://www.nature.com/articles/s41746-018-0029-1

Rajpurkar, Pranav, Awni Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Ng, "Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks," *Cornell University arXiv*, 2017. https://arxiv.org/pdf/1707.01836.pdf

Rama Fiorini, Sandro, Bermejo-Alonso, et al., "A Suite of Ontologies for Robotics and Automation", *IEEE Robotics & Automation Magazine*, March 2017.

Raman, Ritu & Bashir, Rashid, "Biomimicry, Biofabrication, and Biohybrid Systems: The Emergence and Evolution of Biological Design", *Advanced Healthcare Materials*, 2017.

Rawls, J. (1971), A Theory of Justice, Cambridge, MA, Harvard University Press.

Reiter, Ehud, and Robert Dale, *Building natural language generation systems,* Cambridge university press, 2000.

Reiter, Raymond, *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems*, Cambridge, Massachusetts: The MIT Press, 2001.

Reuters, "The Rise of Lousy and Lovely Jobs", 13 April 2012. https://www.reuters.com/article/idUS380409786120120412

Reynolds, Carson, and Picard, Rosalind, "Affective Sensors, Privacy, and Ethical Contracts," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI 2004 Extended Abstracts on Human Factors in Computing Systems, pp. 1103–1106, Vienna, Austria, ACM, 2004.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin, "Semantically equivalent adversarial rules for debugging nlp models," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 856-865, 2018., p. 856

Rickenberg, Raoul, and Byron Reeves, "The effects of animated characters on anxiety, task performance, and evaluations of user interfaces," *CHI '00: Proceedings of the SIGCHI conference on Human factors in computing systems*, New York, NY, USA: ACM Press, 2000, pp. 49–56.

Rid, Thomas, "Cyberwar – does it exist?", *Nato Review Magazine*. https://www.nato.int/docu/review/2013/Cyber/Cyberwar-does-it-exist/EN/index.htm

Riek, Laurel & Howard, Don, "A Code of Ethics for the Human-Robot Interaction Profession", *We Robot*, 2014.

Riek, Laurel D., Rabinowitch, Tal-Chen, Chakrabarti, Bhismadev, & Robinson, Peter, "How Anthropomoprhism Affects Empathy Towards Robots", Cambridge, 2009.

Riek, Laurel, and Don Howard. "A code of ethics for the human-robot interaction profession." *Proceedings of We Robot* (2014).

Rigby, Michael J., "Ethical Dimensions of Using Artificial Intelligence in Health Care," *AMA Journal of Ethics*, Vol. 21, No. 2, 2019, p. 122.

Righetti, L., Q.-C. Pham, R. Madhavan, and R. Chatila, "Lethal Autonomous Weapon Systems", *IEEE Robotics & Automation Magazine*, March 2018, p. 124.

Robert Kitchin, (2014). 'Big Data, new epistemologies and paradigm shifts' in *Big Data & Society*, 1(1): 1-12.

Robertson, Jennifer, *Robo Sapiens Japanicus: Robots, Gender, Family, and The Japanese Nation,* University of California Press, 2017.

Robitzski, Dan, "New AI Generates Horrifyingly Plausible Fake News," *Futurism*, May 30, 2019. https://futurism.com/ai-generates-fake-news.

RobotWorx, "How Can Industrial Robots Improve My Profits?" Accessed December 2018. robots.com/faq/how-can-industrial-robots-improve-my-profits

Rodrigues, Rowena and Anais Resseguier, "The underdog in the AI ethical and legal debate: human autonomy", *Ethics Dialogues*, 12 June 2019. https://www.ethicsdialogues.eu/2019/06/12/the-underdog-in-the-ai-ethical-and-legal-debate-human-autonomy/

Rodrigues, Rowena and Jansen, Philip, "Brief report of the SIENNA foresight workshop on the social and ethical issues of AI and robotics", SIENNA, January 2019.

Rodrigues, Rowena, et. al., *D1.1: The consortium's methodological handbook*, WP1, 2018, Public deliverable report from the SIENNA project.

Roff quoted in Lachow, "The Upside and Downside of Swarming Drones," *Bulletin of the Atomic Scientists*, Vol. 73, No. 2, 2017, p. 96

Roff, Heather and Richard Moyes, "Meaningful Human Control, Artificial Intelligence and Autonomous Weapons" Briefing paper prepared for the Informal Meeting of Experts on Lethal Autonomous Weapons Systems, UN Convention on Certain Conventional Weapons, April 2016.

Romano, Aja, "Jordan Peele's simulated Obama PSA is a double-edged warning against fake news," *Vox*, April 18, 2018. https://www.vox.com/2018/4/18/17252410/jordan-peele-obama-deepfake-buzzfeed

Romano, Donato, Donati, Elisa, Benelli, Giovanni & Stefanini, Cesare, "A Review on Animal-Robot Interaction: From Bio-Hybrid Organisms to Mixed Societies", *Biological Cybernetics*, October 2017.

Rossi, Francesca, "Human-AI Collaboration: Technical & Ethical Challenges," *OECD Conference*, October 26, 2017. http://www.oecd.org/going-digital/ai-intelligent-machines-smart-policies/conference-agenda/ai-intelligent-machines-smart-policies-rossi.pdf

Rueben, Matthew, Bernieri, Frank & Grimm, Cindy et al., "Framing Effects on Privacy Concerns about a Home Telepresence Robot", *2017 IEEE International Conference on Human-Robot Interaction*, March 2017.

Russell, S., and Norvig, P., *Artificial Intelligence: A Modern Approach,* 3rd ed., Essex, Pearson, 2016, p. 929.

Ryan, Mark, Philip Brey, Kevin Macnish, Tally Hatzakis, Owen King, Jonne Maas, Ruben Haasjes, Ana Fernandez, Sebastiano Martorana, Isaac Oluoch, Selen Eren, and Roxanne Van der Puil, *Report on Ethical Tensions and Social Impacts.* SHERPA Project, 2019, https://doi.org/10.21253/DMU.8397134.

Sadowski, Jathan, "Exoskeletons in a Disabilities Context: The Need for Social and Ethical Research", *Journal of Responsible Innovation*, May 2014.

Salem, Maha & Dautenhahn, Kerstin, "Evaluating Trust and Safety in HRI: Practical Issues and Ethical Challenges", *The Emerging Policy of Ethics of Human Robot Interaction,* March 2015.

Sami Coll. (2014). Power, knowledge, and the subjects of privacy: understanding privacy as the ally of surveillance. *Information, Communication & Society*, 17(10), 1250-1263; Gordon Hull, (2015). Successful failure: what Foucault can teach us about privacy self-management in a world of Facebook and big data. Ethics and Information Technology, 17(2), 89-101; Omar Tene and

Polonetsky, J. (2012). Big data for all: Privacy and user control in the age of analytics. *Northwestern Journal of Technology & Intellectual Property*, 11(5): 238-273.

Santoni de Sio, Filippo, and Jeroen Van den Hoven, "Meaningful human control over autonomous systems: a philosophical account," *Frontiers in Robotics and AI,* Vol. 5, No. 15, 2018, DOI: 10.3389/frobt.2018.00015.

SATORI, "CEN Workshop Agreement: Ethics assessment for research and innovation - Part 2: Ethical impact assessment framework, CWA 17145-2, June 2017. http://satoriproject.eu/media/CWA17145-23d2017.pdf

Scharre, Paul, "Robotics on the Battlefield Part II. The Coming Swarm", Center for a New American Security, October 2014, p. 5–6; Magnuson, op. cit, 2016.

Schlogl, L., and A. Sumner, A., "The Rise of the Robot Reserve Army: Automation and the Future of Economic Development, Work, and Wages in Developing Countries," *SSRN Electronic Journal*, 2018. doi: 10.2139/ssrn.3208816

Searle, John R., *Mind, Language and Society: Philosophy in the Real World*, Phoenix, New York, 1999.

Selbst, Andrew D., "Disparate Impact in Big Data Policing," *Georgia Law Review*, Vol. 52, Issue 1, 2017, p. 120. Hao, Karen, "AI is sending people to jail – and getting it wrong", *MIT Technology Review*, 21 January 2019. https://www.technologyreview.com/s/612775/algorithms-criminal-justice-ai

Selgelid, M., 'Dual-Use Research Codes of Conduct: Lessons from the Life Sciences', *NanoEthics*, Vol. 3, No. 3, pp. 175-183, 2009.

Shakarian, Paulo, Jana Shakarian, and Andrew Ruel, *Introduction to cyber-warfare: a multidisciplinary approach*, Amsterdam, Morgan Kaufmann Publishers, 2013.

Shamsuddin, Syamimi, Yussof, Hanafiah & Ismail, Luthffi, et al., "Initial Response of Autistic Children in Human-Robot Interaction Therapy with Humanoid Robot NAO", *IEEE*, March 2012.

Shannon Liao, 'Chinese Facial Recognition System Mistakes a Face on a Bus for a Jaywalker', *The Verge*, 22 November 2018 <https://www.theverge.com/2018/11/22/18107885/china-facial-recognition-mistaken-jaywalker> [accessed 31 May 2019].

Shariff, Azim & Rahwan, Iyad, "Psychological Roadblocks to the Adoption of Self-Driving Vehicles", *Nature: Human Behaviour*, September 2017.

Sharkey, Amanda, and Noel Sharkey, "Granny and the robots: ethical issues in robot care for the elderly," *Ethics and Information Technology*, Vol. 14, No. 1, 2010, pp. 27–40.

Sharkey, Noel & Sharkey, Amanda, "The Crying Shame of Robot Nannies: An Ethical Appraisal", *Interaction Studies*, 11(2), 2010.

Shedletsky, Anna-Katrina, "When Factories Have a Choice, It's Best to Start with People", *Forbes*, June 2018.

Sheh, Raymond, and Isaac Monteath, "Defining Explainable AI for Requirements Analysis," *KI-Künstliche Intelligenz,* Vol. 32, No. 4, 2018, pp. 261-266., p. 263

Simon, Matt, "Companion Robots are Here. Just Don't Fall in Love with Them", *Wired*, February 2017.

Simon, Matt, "The Serious Security Problem Looming Over Robotics", *Wired Science*, August 2018.

Simonite, T., 'Machines Taught by Photos Learn a Sexist View of Women', *Wired*, August 21, 2017. https://www.wired.com/story/machines-taught-by-photos-learn-a-sexist-view-of-women/

Simonite, T., 'When It Comes to Gorillas, Google Photos Remains Blind', *Wired*, January 11, 2018. https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

Simpson, Trudy, "Rise of the Healthcare Robots: Five Ethical Issues to Consider", *Christian Medical Fellowship*, March 2016.

Simshaw, Drew, "Ethical Issues in Robo-Lawyering: The Need for Guidance on Developing and Using Artificial Intelligence in the Practice of Law," *Hastings Law Journal*, Vol. 70, 2019, pp. 173–214.

Singler, Beth, "Are We Expecting Automation to Give Us Modern Day Slaves?", *World Economic Forum*, May 2018.

Skrzypczak, Tomasz, Krela, Rafal & Kwiatkowski, Wojciech et al, “Plant Science View on Biohybrid Development”, *Frontiers in Bioengineering and Biotechnology*, 2017.

Smolensky, “Connectionist AI, Symbolic AI, and the Brain”, 97–98.

Snow, Jackie, “A Robot’s Biggest Challenge? Teenage Bullies”, *Technology Review: Intelligent Machines*, March 2018.

Solinsky, Ryan & Specker Sullivan, Laura, “Ethical Issues Surrounding a New Generation of Neuroprostheses for Patients with Spinal Cord Injuries”, *PM&R* 10(9), September 2018.

Solon Barocas and Andrew D. Selbst, ‘Big Data’s Disparate Impact’, *California Law Review*, Vol. 194, 2016, p. 674; Sarah Brayne, ‘Big Data Surveillance: The Case of Policing’, *American Sociological Review*, Vol. 82, No. 5 2017, p. 978.

Soraker, Johnny Hartz, and Philip Brey, "Ambient intelligence and problems with inferring desires from behaviour," *International Review of Information Ethics,* Vol. 8, no. 1, 2007, pp. 7-12.

Sparrow, Robert & Howard, Mark, “When Human Beings are Like Drunk Robots: Driverless Vehicles, Ethics, and the Future of Transport,” *Transportation Research Part C: Emerging Technologies*, July 2017.

Sparrow, Robert, “Robotics Has a Race Problem,” *Science, Technology, & Human Values*, 2019.

Sparrow, Robert, “Robots in aged care: a dystopian future?” *AI & Society*, Vol. 31, No. 4, 2016, pp. 445–454.

Sparrow, Robert, “Robots, Rape, and Representation”, *International of Journal of Social Robotics*, May 2017.

Spilka, Dmytro, “How AI Is Keeping Us Safe From Drivers Who Use Their Mobile Phones at the Wheel,” *Datafloq*, February 15, 2019. https://datafloq.com/read/aikeeping-safe-drivers-using-mobile-phones-wheel/6064.

Springer, A., Garcia-Gathright, J., and Cramer, H., ‘Assessing and Addressing Algorithmic Bias – But Before We Get There’, *2018 AAAI Spring Symposium Series*, March 2018, pp. 450-454. https://www.aaai.org/ocs/index.php/SSS/SSS18/paper/viewPaper/17542.

Stahl, B. C. (2006). Responsible computers? A case for ascribing quasi-responsibility to computers independent of personhood or agency. Ethics and Information Technology, 8(4), 205–213.

Stahl, Bernd Carsten & Coeckelbergh, Mark, “Ethics of Healthcare Robotics: Towards Responsible Research and Innovation”, *Robotics and Autonomous Systems, 86*, December 2016.

Stansbury, Richard, Olds, Joshua & Coyle, Eric, “Ethical Concerns of Unmanned and Autonomous Systems in Engineering Programs”, *121st ASEE Annual Conference & Exposition*, June 2014.

Steinfeld, Aaron, “Ethics and Policy Implications for Inclusive Intelligent Transportation Systems,” *Robotics Institute at Carnegie Mellon University*, 2010.

Stephens, Tim, “Robotics Project Aims to Develop Systems for Human-Robot Collaboration”, *UC Santa Cruz Newscenter*, December 2012.

Stobbs, Nigel, Bagaric, Mirko, and Hunter, Dan, “Can Sentencing Be Enhanced by the Use of Artificial Intelligence?,” *Criminal Law Journal*, Vol. 41, Issue 5, 2017, pp. 261–77.

Stocker, Michael, “Abstract and concrete value: Plurality, conflict and maximization,” in *Incommensurability,,,,Incomparability and Practical Reason*, R. Chang, Ed., Cambridge, MA, USA: Harvard Univ. Press, 1997.

Sullins, John P., “Introduction: Open Questions in Robotics”, *Philosophy & Technology*, Vol. 24, 2011, pp. 233-238.

Sullins, John P., “Introduction: Open Questions in Robotics”, *Philosophy & Technology*, Vol. 24, 2011, pp. 233-238.

Sullivan, Hannah R. and Schweikart, Scott J., “Are Current Tort Liability Doctrines Adequate for Addressing Injury Caused by AI?,” *AMA Journal of Ethics*, Vol. 21, No. 2 , 2019.

Susskind, Richard and Daniel Susskind, *The Future of Professions: How Technology Will Transform the Work of Human Experts*, Oxford University Press, Oxford, 2015.

Suster, Simon, Stephan Tulkens, and Walter Daelemans, "A Short Review of Ethical Challenges in Clinical Natural Language Processing," *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 2017.

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199,* 2013.

Tal Zarsky (2016) The trouble with algorithmic decisions an analytic road map to examine efficiency and fairness in automated and opaque decision making. Science, Technology & Human Values 41(1): 118–132.

Tamburrini, Guglielmo. "Robot Ethics: A View from the Philosophy of Science", *Ethics and robotics*, 2009.

Tarus, Niu, and Mustafa, "Knowledge-Based Recommendation".

Tatman, Rachael, "Gender and Dialect Bias in YouTubes Automatic Captions," *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 2017.

Tejima, N., "An ethical discussion on introducing rehabilitation robots for people with disabilities." *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*, 2009. doi: 10.1109/roman.2009.5326242

Ter Karppi, '"The Computer Said So": On the Ethics, Effectiveness, and Cultural Techniques of Predictive Policing', *Social Media + Society*, 2018, p. 1.

Tesla, 'A Tragic Loss', June 30, 2016. https://www.tesla.com/blog/tragic-loss.

Thaler, Richard H., and Cass R. Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness*, Penguin Books, New York, 2008.

The Boston Consulting Group, "The Shifting Economics of Global Manufacturing", *February 2015*.

Thieltges, Andree, Florian Schmidt, and Simon Hegelich, "The devil's triangle: Ethical considerations on developing bot detection methods," *2016 AAAI Spring Symposium Series*, 2016.

Thornton, Sarah, Pan Selina, Erlien, Stephen & Gerdes, Christian, "Incorporating Ethical Considerations Into Automated Vehicle Control", *IEEE Transactions on Intelligent Transportation Systems* 18(6), June 2017.

Tiffany, Kaitlyn, "A Timeline of High-Profile Tech Apologies.," *Vox*, July 26, 2019. https://www.vox.com/the-goods/2019/7/26/8930765/tech-apologies-former-facebook-google-twitter-employees-list

Todd, Sarah, "Inside the surprisingly sexist world of artificial intelligence," *Quartz*, October 25, 2015. Retrieved at https://qz.com/531257/.

Toiviainen, "Symbolic AI Versus Connectionism in Music Research", 1.

Toon Calders, Kamiran, F and Pechenizkiy, M (2009) Building classifiers with independency constraints. In: Data mining workshops, 2009. ICDMW'09; Cohen IG, Amarasingham R, Shah A, et al. (2014) The legal and ethical concerns that arise from using complex predictive analytics in health care. Health Affairs 33(7): 1139–1147; and Anthony Danna and Gandy OH Jr (2002) All that glitters is not gold: Digging beneath the surface of data mining. Journal of Business Ethics 40(4): 373–38.

Torresen, Jim, "A Review of Future and Ethical Perspectives of Robotics and AI", *Frontiers in Robotics and AI: Evolutionary Robots*, January 2018.

Trentelman, Kerry, *Survey of Knowledge Representation and Reasoning Systems*, Defence Science and Technology Organisation, Edinburgh, S. Aust., 2009. https://apps.dtic.mil/dtic/tr/fulltext/u2/a508761.pdf

Trentesaux, Damien and Raphael Rault, "Designing Ethical Cyber-Physical Industrial Systems," *IFAC PapersOnLine*, Vol. 50, No. 1, 2017.

Trewin, S., "AI Fairness for People with Disabilities: Point of View," arXiv preprint arXiv:1811.10670, 2018.

Trimmer, Jelte, Pel, Bonno & Kool, Linda, et al., "5.3 Big Data", *Converging Roads: Linking Self-Driving Cars to Public Goals,* February 2015.

Turkle, Sherry, "Authenticity in the age of digital companions," *Interaction Studies*, Vol. 8, No. 3, 2007, pp. 501–517.

Turkle, Sherry, *Alone Together: Why We Expect More from Technology and Less from Each Other*, Basic Books, 2011.

Tutt, Andrew, "An FDA for Algorithms," *SSRN Electronic Journal*, 2016.

Tzafestas, Spyros, "Ethics and law in the internet of things world," *Smart Cities*, Vol. 1, no. 1, 2018, pp. 98-120., pp. 112-115

UCLA Smueli Newsroom, "UCLA Bioengineering Leads Development of Stingray-Inspired Soft Biobot", January 2018.

UK Government, "A guide to using artificial intelligence in the public sector", 10 June 2019. https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector

UN General Assembly, "Universal declaration of human rights," *UN General Assembly* 302.2, 10 December 1948, 217 A (III). https://www.ohchr.org/EN/UDHR/Documents/UDHR_Translations/eng.pdf.

UNESCO, "Section 3.4.4 Companion Robots", *Report of Comest on Robotics Ethics*, September 2017.

UNIDIR, "The Weaponization of Increasingly Autonomous Technologies: Autonomous Weapon Systems and Cyber Operations", *UNIDIR Resources*, No. 7, 2017, p. 2.

UNIDIR, op. cit., 2017 and UNIDIR, "The Weaponization of Increasingly Autonomous Technologies: Artificial Intelligence", *UNIDIR Resources*, No. 8, 2018.

Union of Concerned Scientists, "Industrial Agriculture" https://www.ucsusa.org/our-work/food-agriculture/our-failing-food-system/industrial-agriculture"

Urbi, J., 'Some Transgender Drivers Are Being Kicked Off Uber's App', CNBC, August 13, 2018. https://www.cnbc.com/2018/08/08/transgender-uber-driver-suspended-tech-oversight-facial-recognition.html

Uskov, Vladimir L., Robert J. Howlett, and Lakhmi C. Jain, (eds.), *Smart education and smart e-learning*. Vol. 41, Springer, 2015.; Utermohlen, "4 Ways AI Is Changing the Education Industry," *Medium*, Towards Data Science, April 12, 2018. https://towardsdatascience.com/4-ways-ai-is-changing-the-education-industry-b473c5d2c706

Van den Hoven, J., & Rooksby, E. (2008). Distributive Justice and the Value of Information: A (Broadly) Rawlsian Approach. In J. van den Hoven, & J. Weckert (eds.), *Information Technology and Moral Philosophy*, Cambridge: Cambridge University Press, 376-396.

Van den Hoven, J., Vermaas, P. & Van de Poel, I. (Eds.), *Handbook of Ethics, Values, and Technological Design. Sources, Theory, Values and Application Domains*. Dordrecht: Springer.

Van Wynsberghe, Aimee & Donhauser, Justin, "The Dawning of the Ethics of Environmental Robots", *Science and Engineering Ethics 24*(6), December 2018.

Van Wynsberghe, Aimee, "Service Robots, Care Ethics, and Design", *Ethics and Information Technology 18*(4), December 2016.

Van Wynsberghe, Aimee, *Healthcare Robots*, 2015.

Van Zoonen, Liesbet, "Privacy concerns in smart cities," *Government Information Quarterly,* Vol. 33, No. 3, 2016, pp. 472-480.

Vanderelst, Dieter, and Alan Winfield, "The Dark Side of Ethical Robots," Cornell University arXiv.org, 2016. https://arxiv.org/abs/1606.02583

Vardi, Moshe, "What the Industrial Revolution Really Tells Us About the Future of Automation and Work", *The Conversation*, September 2017.

Vassallo, T., E. Levy, M. Madansky, H. Mickell, B. Porter, M. Leas, and J. Oberweis, *Elephant in the Valley*, 2015. Retrieved at https://www.elephantinthevalley.com/.

Vayena, Effy, et al., “Ethical Challenges of Big Data in Public Health”, *PLoS Comput Biol.*, Vol. 11, No. 2, 2015.

Veale, M., and Binns, R., ‘Fairer Machine Learning in the Real World: Mitigating Discrimination Without Collecting Sensitive Data’, *Big Data & Society*, Vol. 4, No. 2, July-December 2017.

Vega, Julio & Canas, Jose, “PiBot: An Open Low-Cost Robotic Platform with Camera for STEM Education”, *Electronics,* December 2018.

Verisk Maplecroft, *Human Rights Outlook 2018*, July 2018.

Veruggio, Gianmarco & Operto, Fiorella, “Roboethics: Social and Ethical Implications of Robotics”, *Springer Handbook of Robotics*, 2008.

Vincent, James, “AI and Robots will Destroy Fewer Jobs than Previous Feared, Says New OECD Report: But the Impact Will Still be Significant, Increasing Societal Division Between the Rich and the Poor”, *The Verge*, April 2018.

Vincent, James, “First Click: Robots will make it even harder for poor countries to get rich,” *The Verge*, March 10, 2016. https://www.theverge.com/2016/5/10/11648062/first-click-robots-will-make-it-even-harder-for-poor-countries-to-get

Vincent, James, “Hollywood Is Quietly Using AI to Help Decide Which Movies to Make,” *The Verge*, May 28, 2019. https://www.theverge.com/2019/5/28/18637135/hollywood-ai-film-decision-script-analysis-data-machine-learning

Vincent, James, “Robots and AI are going to make social inequality even worse, says new report,” *The Verge*, July 13, 2017. https://www.theverge.com/2017/7/13/15963710/robots-ai-inequality-social-mobility-study

Vollmer, A., R. Read, D. Trippas, and T. Belpaeme, “Children conform, adults resist: A robot group induced peer pressure on normative social conformity,” *Science Robotics*, Vol. 3, No. 2, 2018.

Wachsmuth, I., “Robots Like Me: Challenges and Ethical Issues in Aged Care,” *Frontiers in Psychology*, Vol. 9, 2018. doi: 10.3389/fpsyg.2018.00432

Wade, Michael, “Psychographics: the Behavioural Analysis That Helped Cambridge Analytica Know Voters' Minds,” *The Conversation*, March 21, 2018. http://theconversation.com/psychographics-the-behavioural-analysis-that-helped-cambridge-analytica-know-voters-minds-93675.

Wagner, A. R., Borenstein, J., and Howard, A., ‘Overtrust in the Robotic Age’, *Communications of the ACM,* Vol. 61, No. 9, pp. 22-24, 2018.

Wall-ye, “MYCE_Vigne”, accessed December 2018. wall-ye.com/index-2.html

Wallach and Allen, “Android Ethics: Bottom-up and Top-down Approaches for Modeling Human Moral Faculties”, 149.

Wamsley, Laurel, “Should Self-Driving Cars Have Ethics?”, *Technology*, October 2018.

Webster-Wood, Victoria, “Biohybrid Robots Built from Living Tissue Start to Take Shape”, *The Conversation*, August 2016.

Webster-Wood, Victoria, Akkus, Ozan & Gurkan, Umut et al. “Organismal Engineering: Toward a Robotic Taxonomic Key for Devices Using Organic Materials”, *Science Robotics Review*, November 2017.

Weiser as cited in Spiekermann, Sarah, and Frank Pallas, "Technology paternalism–wider implications of ubiquitous computing," *Poiesis & praxis,* Vol. 4, No. 1, 2006, pp. 6-18., p. 7

Wellcome Trust and Future Advocacy, “Ethical, Social, and Political Challenges of Artificial Intelligence in Health”, Wellcome Trust, April 2018, pp. 12–13.

West, Darrell M., “What Happens if Robots Take the Jobs? The Impact of Emerging Technologies on Employment and Public Policy”, *Centre for Technology Innovation at Brookings,* October 2015.

West, S. M., Whittaker, M., and Crawford, K., ‘Discriminating Systems: Gender, Race, and Power in AI’, 2019, p. 5. https://ainowinstitute.org/discriminatingsystems.html.

Williams, Randi, Christian Vázquez Machado, Stefania Druga, Cynthia Breazeal, and Pattie Maes, “"My doll says it's ok": a study of children's conformity to a talking doll.” In *Proceedings of the*

17th ACM Conference on Interaction Design and Children (IDC '18). ACM, New York, NY, USA, 625-631.

Wilson, Dennis G, “The Ethics of Automated Behavioral Microtargeting,” *AI Matters,* Vol. 3, No. 3, 2017, pp. 56-64.; Jacobson, Jenna, Anatoliy Gruzd, and Ángel Hernández-García, “Social media marketing: Who is watching the watchers?,” *Journal of Retailing and Consumer Services*, available online March 20, 2019, in press. https://doi.org/10.1016/j.jretconser.2019.03.001.

Wilson, Richard L., “Ethical Issues with Use of Drone Aircraft”, IEEE International Symposium on Ethics in Science, Technology and Engineering, May 2014.

Wiseman, Sam, and Alexander M. Rush, “Sequence-to-Sequence Learning as Beam-Search Optimization,” *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.

Wolf, M. J., Miller, K., and Grodzinsky, F. S., ‘Why We Should Have Seen That Coming: Comments on Microsoft’s Tay “Experiment”, and Wider Implications”, *ACM SIGCAS Computers and Society*, Vol. 47, No. 3, pp. 54-64, September 2017.

World Economic Forum, *The Global Gender Gap Report 2018.* Retrieved at http://www3.weforum.org/docs/WEF_GGGR_2018.pdf.

World Economic Forum, *Towards a Reskilling Revolution. A Future of Jobs for All*, 2018b, http://www3.weforum.org/docs/WEF_FOW_Reskilling_Revolution.pdf.

Worms, Frédéric, “The Two Concepts of Care. Life, Medicine, and Moral Relations.” *Esprit*, No. 1, Jan 2006, pp. 141-156.

Wright Scott A., and Schultz, Ainslie E., “The Rising Tide of Artificial Intelligence and Business Automation: Developing an Ethical Framework,” *Business Horizons*, Vol. 61, 2018, p. 824.

Wright, David, “The Dark Side of Ambient Intelligence,” *Info,* Vol. 7, No. 6, 2005, pp. 33–51., p. 34; Brey, Philip, “Freedom and Privacy in Ambient Intelligence,” *Ethics and Information Technology*Vol. 7, No. 3, 2005, pp. 157–166., p. 4

Wu, S., Wieland, Farivar, O., and Schiller, J., ‘Automatic Alt-text: Computer-generated Image Descriptions for Blind Users on a Social Network Service’, *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, Portland, Oregon, USA, pp. 1180-1192, February 25-March 1, 2017.

Yana Welinder, “Police Robots Could Reduce the Use of Deadly Force”, *The New York Times*, 14 July 2016. https://www.nytimes.com/roomfordebate/2016/07/14/what-ethics-should-guide-the-use-of-robots-in-policing/police-robots-could-reduce-the-use-of-deadly-force

Yeoman, Ruth, “Can Artificial Intelligence Give Our Lives Meaning?,” AIA News, Issue 69, June 26, 2018. https://iai.tv/articles/can-ai-generate-meaning-in-our-lives-auid-1101

Zaleski, Andrew, “Man and Machine: The New Collaborative Workplace of the Future”, *CNBC Tech*, October 2016.

Zerilli, J., Knott, A., Maclaurin, J., and Gavaghan, C., ‘Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?’, *Philosophy & Technology*, 2018, https://doi.org/10.1007/s13347-018-0330-6.

Zerilli, John and Gavaghan, Colin, “Call for Independent Watchdog to Monitor NZ Government Use of Artificial Intelligence,” *The Conversation*, 27 May 2019.

Zerilli, John, Alistair Knott, James Maclaurin, and Colin Gavaghan, "Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?," *Philosophy & Technology*, 2018, pp. 1-23.

Zhang, Chuang, Wang, Wenxue & Xi, Ning, et al., “Development and Future Challenges of Bio-Syncretic Robots”, *Research Robotics Review*, August 2018.

Zheng, Wujie, Wenyu Wang, Dian Liu, Changrong Zhang, Qinsong Zeng, Yuetang Deng, Wei Yang, Pinjia He, and Tao Xie, "Testing untestable neural machine translation: An industrial case," *Proc. 41st International Conference on Software Engineering: Companion, Poster*, 2019.

Zimmerman, Terry, and Subbarao Kambhampati, "Learning-assisted automated planning: looking back, taking stock, going forward," *AI Magazine,* Vol. 24, no. 2, 2003, pp. 73-73.