University of Plymouth

PEARL **https://pearl.plymouth.ac.uk**

04 University of Plymouth Research Theses 01 Research Theses Main Collection

2019

Challenging Robot Morality: An Ethical Debate on Humanoid Companions, Dataveillance, and Algorithms

Stamboliev, Eugenia

http://hdl.handle.net/10026.1/14295

http://dx.doi.org/10.24382/389
University of Plymouth

All content in PEARL is protected by copyright law. Author manuscripts are made available in accordance with publisher policies. Please cite only the published version using the details provided on the item record or document. In the absence of an open licence (e.g. Creative Commons), permissions for further reuse of content should be sought from the publisher or author.

Challenging Robot Morality

An Ethical Debate on Humanoid Companions, Dataveillance, and Algorithms

by

Eugenia Viktoria Stamboliev

A thesis submitted to the University of Plymouth in a partial fulfilment of the degree

DOCTOR OF PHILOSOPHY

School of Art, Design, and Architecture
University of Plymouth

June 2019

COPYRIGHT STATEMENT

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without the author's prior consent.

Acknowledgements

I would not have considered beginning this PhD adventure without the academic support and intellectual insights of Georgio Agamben, Pierre Alferi, Anne Dufourmantelle, Catherine Malabou, and Siegfried Zielinski. A huge thanks goes to each of them.

I thank my PhD supervisors Martha Blassnigg, whose absence still remains absurd and tragic, Hannah Drayson, and Michael Punt – the cornerstones of this PhD endeavour – who offered inspiration, questions, and challenges, through which I have grown and developed a coherent piece of work, hopefully. This PhD project grounds on the huge efforts of Sue Denham, Martha Blassnigg and Michael Punt (and others) who manifested this unique research opportunity as part of CogNovo.

I am hugely appreciative of the dedicated feedback that I received from my examiners, Min Wild and Jan Beatens, which improved this thesis immensely. A special thanks goes to Sana Murrani and Mona Nasser, whose wider trust and support mean a great deal. Confronted with the endless admin and deadlines around a PhD, I thank Mandy McDonald for assisting that kindly. I am very grateful to Jane Hutchinson and Susan Waterer for their precious time and support during the last few steps regarding editing.

I owe my colleagues and friends in Berlin, Saas-Fee and Plymouth - and those spread all over the world - so much more than gratitude. There is no personal and intellectual growth without inspiring and amazing others.

Meinen FreundInnen, die diese Zeilen lesen könnten, danke ich für ihre Liebe, Gedanken und Zuspruch, von nah und fern. Ihr seid meine essentiellen Pfeiler und meine Heimat, wohin ich auch gehe.

Благодаря на родните ми, който винаги ми липсват, дори и след всички тези години раздяла.

Благодаря на любимите ми родители за всичко което направиха за мен. Дължа много на вашите грижи и жертви, вашата подкрепа и любов.

Tim, thanking you cannot imply as much as it should. This *monster* became only manageable through your emotional, intellectual and practical support, and your ‘thesis muffins’ baked with your boundless belief in me.

Author's Declaration

At no time during the registration for the degree of *Doctor of Philosophy* has the author been registered for any other University award without prior agreement of the Doctoral College Quality Work submitted for this research degree at the University of Plymouth has not formed part of any other degree either at the University of Plymouth or at another establishment.

This study was financed with the aid of a studentship from the University of Plymouth as part of the CogNovo program and in collaboration with Transtechnology Research.

Word count: 69.986

Date: 7 June 2019

Signature:

Publications:

Stamboliev, E., (2019). Münsterberg, Flusser and the Screen Bodies [Book chapter]. In: Steinmetz, R., ed., *A Treasure trove. Friend of the Photoplay – Visionary – Spy? New trans-disciplinary Approaches to Hugo Münsterberg's Life and Oeuvre*. Leipzig: Leipzig University Press, pp. 109-129.

Stamboliev, E. (2017). 'On 'Spillikin. A Love Story, or on the issues around technology performing sociability' [peer-reviewed journal article]. Off the Lip: Collaborative Approaches to Cognitive Innovation. In: *Avant*, 8 (Special Issue), pp. 265-271. DOI: 10.26913/80s02017.0111.0024

Stamboliev, E. (2017) 'Visualising the Ordinary. On Pierre Schoeller's 'The Lost Time' [peer-reviewed journal article]. *Unmediated. Journal for Politics and Communication,* 1. pp. 77-80.

Stamboliev, E., Jackson, A. (2017). 'Human devices - Instrumentation of physicality in performance'. In *Transtechnology Research Reader 2014/2015*. Plymouth:

Transtechnology Research Institute, pp. 92-101.

Stamboliev, E., (2016) Vilém Flusser - Without firm ground [review: Vilém Flusser exhibition *Without Firm Ground*. Academy of Arts Berlin]. Leonardo Reviews [Online]. Available at http://leonardo.info/reviews/feb2016/exhibition-bodenlos.php

Stamboliev, E., (2015) Telepathy, Hypnosis and the Medium [Conference paper]. In: Punt, M., Denham, S., ed., *Off the Lip. Transdisciplinary Approaches to Cognitive Innovation*. 7-11. September 2015. Plymouth: Plymouth University, pp. 205-214.

Stamboliev, E. (2015). Shabbat [Encyclopaedia entry]. In: Zielinski, S., Weibel P., and Irrgang, D., ed., *Flusseriana. An Intellectual Toolbox*. Minneapolis: Univocal Publishing.

Stamboliev, E. (2015). Dwelling [Encyclopaedia entry]. In: Zielinski, S., Weibel P., and Irrgang, D., ed., *Flusseriana. An Intellectual Toolbox*. Minneapolis: Univocal Publishing.

Stamboliev, E. (2015). On Wittmann, On Münsterberg, On Cinema. In: Punt, M., Blassnigg, M., Drayson, H., Woodward, M., ed., *Transtechnology Research Reader 2014/2015*. Plymouth: Transtechnology Research Institute.

Presentations at conferences:

Stamboliev, E. (2018), "The caring gaze. Simulating, mediating and rethinking care algorithmically' (Paper), Ideoplasticity and the problem of felicitous falsehoods. Transtechnology Research Slow Conference, Plymouth, UK.

Stamboliev, E. (2017), "The Refugee as Image, and Image Maker. Participatory Filmmaking as Journalistic Intervention" (Paper), Future Imperfect symposium, University of Plymouth, Plymouth, UK.

Stamboliev, E. (2017), "Documentary film as a method to ground the *Ordinary*. On Pierre Schoeller's Attempt to Visualise Daily Life in 'The Lost Time'" (Paper), On the Move. Migration, Societies and Change, Manchester University, Manchester, UK.

Stamboliev, E. (2017), "Performing Emotions: Humanoid Robots Beyond Bad Acting" (Paper), *Creative Encounters with Science and Technology. Legacies, Imaginaries and Futures*. Kochi-Muziris Biennale 2016, Cochin, India.

Stamboliev, E. (2016), "Using the Refugee as Vehicle for Political Shift. Unpacking the Austrian Presidential Campaign 2016" (Paper), Media and Migration conference, ECREA - European Communication Research and Education Association, Prague, CZ.

Stamboliev, E. (2016), "A Space to Wonder – Collective Improvisation with Sound and Movement' (workshop in cooperation with Klara Łucznik, Abigail Jackson and Ali Northcott), *Off the Lip 2016. Transdisciplinary Approaches to Cognitive Innovation* conference, University of Plymouth, Plymouth, UK.

Stamboliev, E. (2016), "In Awe of Things. On the Authority of Objects in Scientific and Theatrical Performances" (Paper), Annual TaPRA conference, Bristol University, Bristol, UK.

Stamboliev, E. (2016), "Münsterberg, Flusser and the Screen Bodies" (Paper), *A Hundred Years of Film Theory. Münsterberg and Beyond: Concepts, Applications, Perspectives*, Leipzig University, Leipzig, Germany.

Stamboliev, E. (2015), "Telepathy, Hypnosis and the Medium" (Paper), *Off the Lip 2015. Transdisciplinary Approaches to Cognitive Innovation*. University of Plymouth, Plymouth, UK.

Abstract

Challenging Robot Morality: An Ethical Debate on Humanoid Companions, Dataveillance, and Algorithms

By Eugenia Stamboliev

In this thesis, I reflect on ethical, moral, and agenthood debates around social and humanoid robots in two ways. I focus on how the technological agency of social robots is understood in ethical canons by shifting from moral concerns in Robot Ethics to data-related ethical concerns in Media and Surveillance Studies. I then move to wider debates on morality, agenthood, and agencies in Machine and Computer Ethics research, so as to highlight that social robots, other robots, machines, and algorithmic structures are often moralised but not understood ethically. In that vein, I distinguish between these two terms to point to a wider critique on the anthropocentric and anthropomorphic tendency in ethical streams, so as to view technology from a morality-aligned standpoint.

I undertake a critical survey of current ethical streams and, by doing so, I establish a transdisciplinary ethical discussion around social robots and algorithmic agencies. I undertake this research in two steps. First, I look at the use of humanoid social robots in elderly care, as discussed in Robot Ethics, and expand it with a view from Media and Surveillance Studies on data concern around robots. I hereby examine the *social robot* and the allocation of its ethical and moral agency as an anthropomorphised and humanoid companion, data tracking device, and Posthumanist ethical network of agencies. This is done to amplify the ethical concerns around its pseudo-agenthood and its potential position as dataveillance. Next, I move on to streams in the Philosophy of Technology (POT) and Machine/Computer Ethics. Here, I discuss concepts on machinic moral agency in digital systems. As I pass from the social robot as a humanoid pseudo-agent towards moralised algorithmic structures, I lay out wider conflicts in morality research streams. Specifically, I address their epistemological simplification and reduction of moral norms to digital code, as well as the increasing dissolvement of accountable agenthood within algorithmic systems.

By creating a transdisciplinary investigation on techno-ethical and techno-moral canons and their agency models, I urge for a holistic ethics that, first, gives a greater focus to human agent accountability and moral concerns in the application of robots and, second, negotiates new moral or social norms around the use of robots or digital media structures. This is aligned with increasing concerns around the growing commodification of health data and the lack of transparency on data ownership and privacy infringement.

List of Contents

Key Terms and Themes

Algorithm

I refer to algorithms and algorithmic structures when more broadly explaining the computational architecture, tracking, and ethical structures of social robots in digital information systems. I do not dwell on this term in detail and need it only to distinguish computational agency from anthropomorphic agency. Hence, the definition found in the *Encyclopaedia of Mathematics* is sufficient for this thesis: 'A computational algorithm is realised in the form of a computational process (i.e. as a finite sequence of states of a real computer, discretely distributed in time).'[1] Introna & Wood's (2004) definition is equally useful, of algorithms as a 'mathematical, or logical, term for a set of instructions' (180).

Algorithmic Surveillance

According to Introna & Woods (2004), the term 'algorithmic surveillance' was coined by Norris and Armstrong (1999) in their pioneering book, *The Maximum Surveillance Society*. They write that surveillance uses automatic step-by-step instructions and specifically refer to surveillance technologies

> 'that make use of computer systems to provide more than the raw data observed. This can range from systems that classify and store simple data, through more complex systems that compare the captured data to other data and provide matches to systems that attempt to predict events based on the captured data' (Introna & Woods, 2004: 181).

Anthropocentrism

This thesis operates with two similar terms, anthropomorphism and anthropocentrism. The Merriam Webster Dictionary defines *anthropocentrism* as 'considering human beings as the most significant entity of the universe' and as 'interpreting or regarding the world in terms of human values and experiences'.[2] This concept creates problems when discussing morality and ethics around robots and machines on various levels. This dynamic unfolds as problematic in wider POT (see Philosophy of Technology)

---

[1] Available at https://www.encyclopediaofmath.org/index.php/Computational_algorithm (Accessed 22.05.2018).

[2] Available at https://www.merriam-webster.com/dictionary/anthropocentric (Accessed 15.05.2018).

discussions, since their ontological views on technology mix with the focus on human consequences and expectations around technology. What the robot is capable of and what it is supposed to understand or intend, then fuses with the ambition to recreate human agency. Floridi (2014) describes the issues around anthropocentric ambitions as linked to thinking agenthood as a singular entity. He denies individual views on technology, but holds on to moralising its capacities. For Posthumanists, anthropocentric views are inherently flawed when thinking or talking about technological agency, not only because agency is not a separate entity, but because some agencies are not human. This theme is discussed in Chapter Four and Seven, and is picked up with the labels of (2) Reductionist Morality and (3) Distributed Morality.

Anthropomorphism

Anthropomorphism implies 'the tendency to attribute human characteristics to inanimate objects, animals or others with a view to helping us rationalise their actions' (Duffy, 2003: 180), and is derived from the Greek words *anthopos*, which means human, and *morphe* meaning form/design. Duffy points out that anthropomorphism favours the observer's perspective in the interaction between robot and human. Social robots afford an anthropomorphic projection as much as *automata*[3] did in the 18th century (Reilly, 2011).

However, this projection process becomes problematic if moral abilities or responsibility are projected into robots, even though these do not embed any human agenthood, human agency or any moral understanding. Posthumanists like Braidotti (2006) do not support an anthropomorphic ethics or agency of anything, but she acknowledges that anthropomorphising technology is an automatised, biological response of human agents and, hence, cannot simply be switched off.

The difference between the anthropomorphic and the anthropocentric - both morality led - approaches is that the anthropomorphic camp is concerned with how the robot is perceived as an agent or tool, while the anthropocentric camp aims to embed moral

[3] 'The word automaton comes from the Greek automatos, meaning 'acting of itself', referring to automated moving figures of animals or human beings. While automata look like dolls or toys, it is their animation that signifies life. This life-like movement means that automata are often perceived as if they're alive. As a result, automata are central to debates about mimesis or the representation of reality in the historical period in which they exist' (Reilly, 2011: 1).

abilities and rules into robots. Both concepts think agenthood, agency or decision-making are dependent on human evaluation and reference systems, which are rarely specified. Anthropomorphism dominates Part One and the discussion on social robots, as well as the (1) Apparent Morality and Agenthood discussion. It is discussed in detail in Chapter Six.

Computational Interaction / Computational Architecture

Computational (or algorithmic) interaction refers to how the software level of robots influences their interactivity and responsiveness. Computational interaction can be understood from various disciplines, but plays a big part in research on the Human-Robot-Interaction (HRI). In this thesis, I explore this theme as aligned to discussions on algorithmic architectures (Rossini, 2012), tracking (Haritaoglu et al., 2000), and algorithmic autonomy (Floridi, 2014; Wallach & Allen, 2009). This theme is debated in Chapter Six in detail.

This theme is important to understand for the agency debate, since social robots are not only humanoid bodies and companions, but embed and operate computational 'capabilities of conducting a wide range of social functions [e.g., speech recognition, speech generation, visual recognition, affective responses, turn-taking (interactivity), and artificial intelligence]' (Lee et al., 2005: 540). Their computational interactivity allows me to move from a discussion on social robots as perceived human-like pseudo-agents, to them as computational systems and algorithmic agencies. Furthermore, it allows me to argue that the ethical concerns can be situated in their 'inner shell' (Read, 2014), not only in their appearance.

Dataveillance

Dataveillance is the 'systematic use of personal data systems in the monitoring or investigation of the actions or communications of one or more persons' (Clarke, 1988). Dataveillance marks a shift from strategic surveillance to the appropriating of data collections, on the grounds of an increase in collected data sets and the expansion of digital information structures (Püschel, 2014; Andrejevic, 2012). I align this concept to social robots, since I fear that the ignorance of Robot Ethics towards data-related concerns (which are masked by having an anthropomorphic discussion on why robots are good or bad companions) neglects their ability to track and collect data, which is

problematic in practical terms. Dataveillance links with the computational interaction of robots and their tracking capacities, but refers to the implementation and expansion of technology into environments such as elderly care, from which data can be appropriated and commodified (Püschel, 2014). This theme is explored in Chapter Three.

Ethical Gestures According to Agamben (2000)

Agamben's (2000, 2014) work offers an ethical angle on technology, one that does not moralise but is concerned with understanding the reciprocity of human-technology agencies. For Agamben, early cinema is a gestural remediation of socio-cultural and technological discourses. His conceptualisation of early cinema as an 'ethical gesture' (2000) allows me to undertake a speculative investigation, which looks at technology as a network of relational agencies of human and technological origins.

I use this concept to support my view on tracking as the robot's means of remediating human and technological agencies; of making its expression ethical; and allowing for a perspective on tracking, which is not moralising but reflects on the complexity of new agency models that are neither stable nor hierarchical. Agamben looks at how cinema technology has embodied and transformed human gestures within a philosophically and historically complex visitation of what he calls the 'crises of representation' in the early 20th century, which I believe can also be traced in the social robot's anthropomorphic agency projection. This theme is explored in Chapter Three and is picked up again in Chapter Six.

Ethics as Aligned to Philosophy of Technology (POT) and Moral Philosophy

Ethics can be understood as the structure or a vessel of moral norms; it can also be referred to the discipline in which morality is discussed. In most traditional moral philosophies (such as Kantian, Humean, Utilitarianism, etc.), there is no distinction between ethical and moral norms. For Luhmann, ethics can be understood as a 'reflective theory of morality' (Luhmann, 1989: 360). I sympathise with Ward's media-theoretical view (2015), which sees *ethics* as a set of principles that can be a singular or plural concept and can refer to a language or a set of norms. But, he argues that ethics must be something truly dynamic, and can never become dogmatic, since it is a 'human activity' (6). He argues that ethics is not just the disposition to adhere to rules, but also the disposition to critique and improve these rules, since ethics is a process of a 'lived

experience of ethical doubt and plurality of values, and then seeks integration and theoretical understanding' (2015: 7). He explains the difference between meta-ethics and applied ethics with:

> 'Meta-ethics asks three big questions about the nature of ethics: What are we saying when we make an ethical claim? How do we know that what we say is justified? Why does ethics exist in the first place? There are plenty of ethical theories, from descriptivism and intuitionism to realism and relativism. Applied ethics, on the other hand, asks not what we mean by ethical concepts like good or right but what *is* good or right, and how to do what is good or right in certain situations' (2015: 6,7).

Ward's view offers a useful link between moral philosophy and POT, but is not useful to understand robot ethical concerns. I consider it a general definition on ethics worth keeping in mind. However, I do not commit to one ethical school on technology, due to the comparative character of this thesis' investigation, since I highlight the inclusion of Posthumanist ethics, together with Agamben's view on the ethical gesture, as exclusive ethical views on technology, so as to question the moral concerns of Robot Ethics to critique traditional POT discussions. Ethical POT streams are increasingly becoming meta-ethical, and do not deal sufficiently with practical questions on the use of robots.

Ethics as Aligned to Posthumanism

I make use of Posthumanist ethics (mainly through Braidotti, 2006 and Barad, 2003) to counter the view from Robot Ethics on social robots as anthropomorphic companions, and Machine and Computer Ethics' computers and algorithms as moral systems. Anthropocentric ethical streams in Machine and Computer Ethics and POT, which refer to the moral qualities of algorithmic structures, as much as anthropomorphic views on social robots as companions, would be rejected by Posthumanists. Braidotti (2006) is critical towards anthropomorphism, but sees it as an imbedded human response; she rejects anthropocentric views on technology as being epistemologically mistaken. The Posthumanist thinking allows for a better understanding of robots as an entanglement of various human and technological agencies. In particular, the view from Braidotti on ethics as an intertwining of agencies, and on the 'ethical complexities' in socio-technological relations (Braidotti, 2006: 16), contributes to a holistic view on robots as bodies, computational machines, and data collections, which also embed tracking capacities and are much more than moralised companion devices or single

bodies. I align the (ii) Media and Surveillance Studies framework with the Posthumanist perspective on robots to understand data and tracking as ethical networks. I further share the ambition of Braidotti (2006) and Zylinska (2014) towards overcoming moralist universalism and reductionism, as much as the concept of a stable, single or rational subject owning certain moral intentions. Equally, I acknowledge that the Posthumanist ethical views are insufficient to practically debate robot or developer/programmer accountability.

Human-Robot-Interaction (HRI)

Human-Robot-Interaction is a research stream in HR that explores how human agencies interact with robots, or how to equip robots so they can interact with humans. This area has two angles; the human and the robotic angle – both requiring different cues to interact. The human or anthropomorphic perspective asks for social robots to appear natural, believable, and responsive, so as to gain a wider human acceptance. From a robotic perspective, the robot needs to be able to read its environment or human movement through the sensors and modules embedded in its computational system. The better the social robot can interact with its environment, the more meaningful and *social* it is considered, but this requires an operational autonomy in managing and synchronising data-processing and locomotion. My concern is that HRI research underestimates the ethical dimension of the data management it enables, since it lacks the insights from Media and Surveillance Studies. This theme is debated in Chapter Three briefly and in Chapter Six with greater detail.

Media and Surveillance Studies (MSS)

A new branch of Media Studies, aligned with Social Sciences and Surveillance Studies, has fused into a research area referred to as Media and Surveillance Studies (MSS) (Kammerer & Waitz, 2015). What MSS enables in this thesis is to amplify the ethical structure of data collection as an ethically problematic one and as inherent to the use of social robots. MSS is an emerging transdisciplinary research area, in which social structures and institutions, surveillance, and data-production merge with historical and cultural media studies and media theory. Inspired by media theorists like Kittler, contemporary researchers in this area address topics on new aesthetics, digital convergences, new and emerging forms of data-production, and their influence on not

only socio-cultural structures, but also on privacy structures. Some of the prominent voices in the field are Andrejevic (2012), Galloway (2012), Gates (2011), Gitelman (2013) and Magnet (2011). This framework guides Chapter Three.

Moral Agency in Machine and Computer Ethics (Philosophy of Technology/POT)

I specifically survey moral agency models in POT to understand the computerisation and algorithmisation of moral agency into algorithmic code. This process leads to reductionist, distributed, and simplified morality concepts that become too abstract to answer for any accountability questions in applied ethics, but that are well suited to explain the network of agencies constituting digital technology. Moral agency models in Machine and Computer Ethics (Wallach & Allen, 2010; Floridi, 2014; Kroes & Verbeek, 2014) often support an anthropocentric view on robotic or computation systems, which aligns technological autonomy and agency to moral reasoning and moral actions. However, these canons often reject anthropomorphic projections and individual agenthood models (like Floridi does), since they instead align to a collective ethics, as opposed to the companion position in Robot Ethics, which looks at the robot as an (pseudo) individual agent.

Some theories come close to the Posthumanist views on socio-material agency (Brey, 2014). The difference between these techno-philosophical views and Posthumanist ethics is that the distribution of moral agency still moralises technological agency, instead of understanding the intersection of agencies as networks without evaluative intentions. An example is the labelling of algorithms as being ‘good’ or ‘evil’ agents (Floridi, 2014); which is something I find problematic.

Further, this moralising process supports an implied but undiscussed value system. These streams make use of evaluative implications and traditional morality concepts, but are not clear on their reference points. Some discussions align with individual ethics (compared to ethical structures); others refer to Kantian ethics or Utilitarianism (Miori & Herschel, 2017). Kantian ethics would focus on the duty of the individual to act morally, whereby the actions become expressive for moral agency. However, a Utilitarian ethical view ‘examines right or wrong based on the consequences of an act or a rule’ (33). The Utilitarian perspective applies the principle of utility to individual moral actions and the Utilitarian rule applies the principle of utility to moral rules’ (33). Appropriating philosophical backdrops from moral philosophies to suit the

technological complexity is, in my view, flawed and too simplified to be a legitimate and holistically ethical approach.

Moral Agency in Robot Ethics

Moral agency is assigned to social robots in Robot Ethics through an anthropomorphic projection of values and intentions, which comes from their humanoid shape. In the framework of Robot Ethics, I only make use of the anthropomorphic view on social robots and the ethical position of social robots in human environments, such as elderly care. I explore this through researchers such as Royakkers & van Est (2016), Lin et al. (2012), Sharkey & Sharkey (2010), and Turkle (2005, 2011).

The moral agency model of social robots makes use of an anthropomorphic fallacy that social robots are companions, and it projects agency onto the robot on the grounds of its humanoid features. The limited understanding of robots ethically, in my view, denies the Posthumanist position that robots could be a network of other, non-human agencies, which lead to different ethical concerns, such as data or privacy issues. It also denies the view on social robots simply being an ethical technology without first having to be involved in moral concerns. The view on social robots as companions and the drawing of moral agency from their humanoid shape is also critiqued by other POT theorists as being too anthropomorphic and superficial.

I make use of the view from Robot Ethics to discuss the anthropomorphic companion agency first. I then shift to the anthropocentric discussions on machine morality and intelligence, and morality in POT streams, which move from a humanoid entity agency to a computational system agency, but both still refer to robots. Robot Ethics is not clearly separated from wider POT discussions and, as a conglomerate between philosophical streams, certain discussions do overlap.

Philosophy of Technology (POT)

I survey wider Philosophy of Technology canons (Kroes & Verbeek, 2014) with a focus on anthropocentric streams in (iii) Machine and Computer Ethics, which I argue as a sub-division in POT. I enter (iii) Machine and Computer Ethics to offer a more holistic ethical view on robots as machines and not only as companions, and on robotic capacities beyond robotic appearance. I establish this two-fold framework to survey how morality/moral agency is debated in contemporary discussions on technological

and artefactual agency (Johnson & Noorman, 2014), robotic machines (Wallach & Allen, 2009) or algorithmic systems (Floridi, 2014). In a wider sense, I distinguish between morality-led Robot Ethics, morality-led POT streams, and ethical views in Posthumanism and, further, between the anthropomorphic tendencies in Robot Ethics and anthropocentric ones in POT.

My wider critique of the morality-led POT streams is on the reductionist conceptuality of morality and the inability to resolve accountability questions. This shows in a superficial and simplified translation of human-based moral philosophical views on agency to technology and in an increased theoretical disconnection between technological agency or autonomy from accountability. The consequences are that most streams divert from practical questions when applying technology such as robots although these are set out to respond to them.

Social and Humanoid Robotics (HR)

HR and Robot Ethics commonly reinforce the anthropomorphic view on social robots. This view comes from social robots mostly being designed as humanoids and positioned as companions on the grounds of their human-like shape. This is concerning in my view, because it allows to discuss them ethically as if they are human agents, while they are not. While the social robot is, technically speaking, a complex machine, the *social* attribute attached to this kind of robot can mean different things; from referring to its humanoid appearance to reinforcing its humanoid shape. But, it can also mean a robot that is interactive and embeds computational responsiveness in a human environment.

HR is a conglomerate of research from Robotics, Psychology, Mechanics, Engineering and Computer Science, which explores the social robot by studying, developing and realising these socially capable machines, equipping them with a very rich variety of capabilities that allow them to interact with people in natural and intuitive ways. This ranges from the use of natural language, body language, and facial gestures, to more unique ways, such as expression through colours and abstract sounds (Read, 2014: i). HR does not always specify the environments in which robots are used, while Lin et al. (2010) and Royakkers & van Est (2016) discuss the use of robots in professional elderly

care as important applications.[4] What makes social robots special is their humanoid shape, aligned with their highly developed computational interactivity. Hence, the social robot is distinguished, for instance, from assembly or war robots (Royakkers & van Est, 2016) due to their different functions and appearances. Consequently, this leads to different ethical debates on these robots. Chapter Six will discuss the HR framework and social robots in detail.

Tracking / Data Tracking

To *track* means to locate, to trace, to contextualise a position of oneself or oneself within the environment. The visual tracking capacity of social robots is used in this thesis to illustrate that the alignment between computational and moral autonomy is not sufficient to understand how a process like tracking transforms into an ethical network. Tracking is used as a cumulative term for the application of tracking modules, or the tracking process in social robots. Tracking is a computational process (Lee et al., 2005; Fong et al., 2002; Brèthes et al., 2004) that is ontologically bound to collect, manage, and process data. Therefore, it allows for a discussion on dataveillance to emerge in relation to social robots. Tracking modules are used for the Human-Robot-Interaction (HRI) to enable the robot to detect, process, and respond to the human movement. HR and Robot Ethics would understand the tracking process as an operational and neutral undertaking, but I urge to reconsider this simplicity, and to acknowledge that tracking is not only a multidimensional process, but also an ethical one. By focusing on understanding this process, this thesis addresses the complexity of tracking technology as an ethical structure, which reflects on the human values and concepts that shaped it, and defines what concepts such as emotion, interaction, and care mean for a robot. Tracking is further explored in Chapter Two, Three, and Six, but, indirectly, its definition is affected by every discussion on the computational capacities of robots and these having ethical or moral intentionality.

[4] This thesis specifically looks at care robots, which Lin et al. (2010) describe as: 'Personal care and companions: Robots are increasingly used to care for the elderly and children, such as RI-MAN, PaPeRo, and CareBot. PALRO, QRIO, and other edutainment robots mentioned above can also provide companionship' (944).

Introduction

The implementation and use of digital technology, such as social robots, raises moral and ethical concerns (Luppicini, 2009; Bunge, 1997; Royakkers & van Est, 2016), as the growing operational autonomy of digital systems complicates questions on what actions or decisions these devices are accountable for. Particularly problematic are social robots when applied into sensitive environments, such as elderly care. I examine how specific agency models on social robots influence and intersect with wider ethical concerns on digital technology, and how a disciplinary segmentation, ultimately, allows for a neglect of practical consequences in the use of robots in elderly care.

I enter this ethical exploration on moral agency of social robots with the objective to understand the connection between humanoid and material agency models, and between the moral and ethical view on digital technology; not by attempting an ethical techno-materialist view on robots, but instead, by exposing various ethical *blind spots* or neglects around robots, which disregard practical consequences on the grounds of theoretical narrowness. Without advocating for one *correct* agency or ethical canon in which to situate social robots, I aim towards a new transdisciplinary ethical discussion on social robots that aligns various agency models to allow for a better view on accountability questions.

One *blind spot* with which I am concerned is around social robots as data tracking[5] devices. My hypothesis is that this connection is not sufficiently explored nor as fully

[5] Tracking is an interactive and complex process of collecting and managing data, and is implemented into the robot to interact with its environment. It is explained in the Key Terms section. Tracking is negatively connotated in the context of website and online application since it is associated with surveillance or unasked detection of online user behaviour or movement. Robot Ethics does not discuss this association, which I consider neglectful.

understood in Robot Ethics (nor in POT streams). I derive this hypothesis from my early focal point of exploration, the dramatic play, *Spillkin. A Love Story* (2017), which led me to identify an unexplored entanglement between ethics, social robots, elderly care, and social robots being tracking devices (Stamboliev, 2017, 2018). The play exposes and reflects upon the relationship between a social (humanoid) robot named Spillikin that 'assists' a woman called Sally, who is about to lose her memory due to Alzheimer's disease. The robot is positioned as her counterpart, friend, and companion (what Robot Ethics and Humanoid Robotics would want it to be) and is meant to be a humanoid substitute for Sally's deceased husband, who designed it. Despite presenting a future-led and hypothetical scenario, the play illustrates an important disconnection in how social robots are presented by HR and Robot Ethics, and in how I suggest discussing them; between companion and tracking device.[6]

Social robots might be promoted as social devices by HR (or by theatrical plays) because they are shaped as humanoids and resemble human bodies, but, in fact, they are interactive, computational technologies that can visually recognise, manage data, synchronise locomotion, and process instructions (Royakkers & van Est, 2016; Breazeal, 2002). What the play exhibits, by accident rather than by design, is not only why robot companionship presents a development of ethical concern, but that there is a conflict between perceived ethical concerns and ontological ethical concerns.

The dominance in promoting robots as perceived companions mirrors the ethical concerns on social robots in Robot Ethics, which focus on how and why social robots

[6] I highlight that at this moment, considering the technological development of robotic systems, social robots are far from being autonomous, responsive or fully interactive technologies. Hence, this topic is still future-oriented. However, the tracking and data concerns are likelier to manifest than the implementation and acceptance of social robots as companions. This thesis is an exploration of the accompanying ethical discussions on robots and less on their actual practical applicability.

are good or bad companions. However, this focus allows for an *anthropomorphic fallacy* to drive the ethical concerns in the field, one that mistakenly expects robots to *be* caring companions and to have human-like attributes or judgement. This then creates a limited view on social robots, which are only debated as *caring companions*[7] or as deceptive pseudo-agents (discussed in Chapter Six). Supposedly, this perceived companion position does make it easier to promote them in environments such as elderly care, where their presence needs to be seen as friendly and caring company, not as a bulky and indifferent device (Chapter Two will elaborate on this).

As I suggested already, there is more to be concerned with around this bulky device *ethically*, which goes beyond the perceptive or promoted position even if this companion position, as limited as it might be, is reflected upon critically by Robot Ethics (Sharkey & Sharkey, 2010; Royakkers & van Est, 2016). Even if the field does not simply endorse the potential, but instead presents various concerns around this position - as does the play - the way in which this is done ratifies the anthropomorphic fallacy as the dominant ethical focus on robots, while masking another one: their tracking abilities. I assume that the companion position is unable to cover the ethical spectrum of other concerns raised by robots beyond how they are perceived.

This leads to Robot Ethics neglecting the second angle I therefore highlighted initially as a *blind spot*: social robots as a digital data collector in a humanoid body. As digital devices equipped with cameras, these technologies are interactive or responsive devices that collect data. What I suggest further is that the companion position is not only

---

[7] The Merriam Webster Dictionary refers to 'care' as a form of 'strict attentiveness to what one is doing' or as 'attention accompanied by protectiveness and responsibility'. Available https://www.merriam-webster.com/thesaurus/care (Accessed 20.04.2017). The simulation of human care through technology focusses mainly on the simulation of emotions and gestures in humanoid robots.

limited, but also *limiting*. It might be problematic, since it masks an ethical discussion on the digital capacities of robots due to their invisibility and complexity, but does not change the fact that these exist or are ethically problematic, and that a vulnerable group of people is exposed to the complexity of a digital device.

As I mentioned, one reason for this masking neglect could reside in the agency model with which HR and Robot Ethics operate. Here, the machinic or computational side of robots is seen as an ethically irrelevant space, and is discussed as instrumental and functional, but not as a relevant ethical network. In this sense, I advocate that understanding the agency angle attributed to the social robot is crucial to comprehend the associated ethical or moral norms, and vice versa.

This forms the starting point of my thesis, which focusses on these two conflicts that I derive from the previous insights. The first conflict lies in the ethical disconnections around social robots as hybrid agents. The already identified twofold position, between being a companion pseudo-agent and a tracking device, makes it hard to assign them to only one ethical canon (be it in Robot Ethics, applied ethical streams, Machine and Computer Ethics, Posthumanism, or Data Ethics). Depending on what ethical view is taken, the ethical concerns and the agency models form differently, but as these concerns change theoretically, the practical capacities of social robots do not. I will elaborate on how social robots, as companions or as tracking devices, are always ethical technologies that are inherently associated to data concerns, while I trace the neglect of this realisation throughout the thesis.

In my view, the first conflict originates from a two-fold position that social robots embody as a technology. On one hand, social robots are designed as humanoid bodies

and are promoted as companions or assistants for professional care (Royakkers & van Est, 2016). This encourages Robot Ethics to focus on human-centred questions, such as: robots potentially harming, deceiving, not responding to, or isolating patients when used in elderly care (Chapter Six will discuss this further). On the other hand, social robots are interactive and responsive devices that manage and produce data (Royakkers & Van Est, 2016; Lee et al., 2005; Fong et al., 2002) and, therefore, I think they present huge challenges for data ethics, patient privacy, and data infringement (Knoppers & Thorogood, 2017; Floridi, 2016; Andrejevic, 2012; Ball et al., 2012). This doesn't seem to concern Robot Ethics as much, given their misconception that robots can be seen as companions *exclusively*. In addition to this two-fold divide, the very application of social robots into elderly care disrupts the traditional work environment and caring norms of professional care – which can also be argued as an *ethical* process worth discussing, but this intrusive process does not find much ethical relevance within Robot Ethics either.

Agency discussions on robots are co-shaped by their ethical discussions, but I identify a second conflict in the conflation of *moral* and *ethical* concerns of social robots, and in the wider moralisation of technology. Robot Ethics and wider Philosophy of Technology (POT) streams do not necessarily distinguish between these two terms, nor do most traditional morality streams.

Furthermore, to the theoretical conflation between ethics and morality in traditional moral philosophy, the concept of morality seems loaded with implied associations of *good* or *bad* virtues, actions or decisions of robots, while it remains often unclear from which specific moral philosophical school these references are taken. My hypothesis in this context is that what POT then attempts – by aligning *morality* to robots or digital

technology – is to associate (non-existing) human values to the operational autonomy of robots. Yet, what they end up with is the reproduction of old moral concerns[8], which are not agreed on in human moral philosophies either. Even less are these translatable into robot technology without facing huge difficulties.

Ethics, on the other side, is a term widely used as a structural vessel, discourse or dynamic in which to discuss moral evaluations and agent positions, while also used to name disciplines like Robot *Ethics*. While morality is always ethical, ethics does not have to be morally bound, since ethical canons can be completely detached from morality as an evaluative concept (see Posthumanist ethics). As I suggest incorporating Posthumanist ethics to understand data concerns of robots better, I identify that the conflation between moral and ethical concerns could become problematic, given that the ethical view from Posthumanism rejects the moralisation of technology (Braidotti, 2006). Posthumanists would, in fact, argue that moral evaluations hinder certain ethical perspectives.

What I attempt is using Posthumanist ethics is to offer a counter concept (aligned with Media and Surveillance Studies) that tackles the moral streams around robots in Robot Ethics and is able to address data concerns more appropriately and holistically. I will incorporate Posthumanist views into my ethical investigation to point out that there are different ethical and conceptual ways to discuss techno-human agencies without requiring a moral evaluation or labelling, and that data concerns are better understood through an inherently ethical discussion on robots, data, and tracking processes. However, this suggestion might prove less suited to resolve or address the practical

[8] For instance, questions on the origins of moral reasoning, or the balance between individual or collective morality or the necessity of individual moral duty in Kantian ethics.

concerns when technological autonomy conflicts with human accountability, and when responsibilities around social robots must be negotiated.

As I advocate to distinguish between ethics and morality, I do not necessarily advocate for one specific moral or ethical canon to be the *correct* one in which to position robots. Instead, I will outline the advantages and disadvantages across anthropomorphic, anthropocentric, and ethical views by reflecting on their agency models of either robots or digital technology, as I move increasingly into POT discussions that leave Robot Ethics behind. In detangling the companion view from the tracking device angle, and in separating moral from ethical concerns, I am sustaining the practical value of this thesis.

This value grounds in social robots embedding computational systems that own, collect, and produce data, which is something that the play analysis pointed out to me; for robots to be *sociable*, this requires them to track movement or emotional expressions and to process data (Stamboliev, 2017). Yet, Robot Ethics does not fully address data issues, even if it incorporates discussions on social robots as monitoring devices that can track data. However, the field has a different approach to agency than, for instance, MSS has, since it does not consider that the very collection of data is always *ethical* and, often unintendedly, *unethical*. I will show that the moralisation of robots often creates a hindrance to understand data concerns fully on the grounds on how their or their developer's agency is understood.

There are specific reasons for taking data concerns seriously, which are linked to the increasing value and commodification of health data (Knoppers & Thorogood, 2017) and the lacking or insufficient media literacy of the elderly - which, in this case, would equip them with the necessary knowledge to understand wider conflicts around data

ownership and privacy. These two concerns cannot be ignored in any research or ethical field, especially since robots are interactive technologies and data-collecting devices from their very moment of being placed into an environment. Stressing this practical consequence is essential, because elderly care is a specifically problematic and sensitive environment in which to place robots, especially before having fully understood or explained their practical application. Hence, the neglect around data-related concerns on social robots makes Robot Ethics *unethical*.

Another motivator to stress data concerns is to start a wider discussion around accountability as crucial to understanding robotic agency and autonomy, as much as human agency is in designing robots and digital systems. My extensive reflection around thinking moral agency in robots will lead me to discussing moral agency in algorithmic systems, raising even more concern around morality and technology beyond social robots as a focal point. One new concern I see relates to morality being an evaluation concept, but, as such, it needs an agent to be assigned as *moral*. For most moral philosophies, moral agency, and agenthood overlap in the human subject. However, as this thesis evolves and new techno-moral streams emerge, the stable human agenthood position dissolves into numeric rules and algorithmic intentionality as it becomes an increasingly abstract entanglement.

This new agency entanglement challenges the alignment of moral agency to robots and algorithms, since what most traditional morality models have in common is: they are supposed to address accountability or responsibility in human-technology interactions. Still, newer canons struggle to do this. Especially, the growing autonomy of robots and algorithmic systems has complicated discussions on accountability, and considering that social robots step into an interactive relationship with humans, more clarity on this

encounter is key. This implies that the operational autonomy of robotic actions, allowing for robots to respond or move, has complicated their ethical position. But has this made them *more* or *less* moral devices? And how has this shifted their ethical dimension? My wider concern is that a lack of theoretical agreement on what actions or behaviours social robots are accountable for means that we cannot justify their application in sensitive contexts, such as elderly care.

To tackle this concern, I offer a theoretical and ethical investigation on agency, agenthood, autonomy, and accountability of robots as humanoid companions, machines and information structures. I will survey, examine, and critique various agency models by outlining their individual strengths and limiting perspectives in accordance to their theoretical and practically relevance. Ultimately, I aim for a transdisciplinary ethics on digital technology or robots; one that urges for a better foundation that does justice to diverse formats of ethics, and also to the agency models on digital technology, recognising their new growing autonomy and ubiquitous applications. In such a new ethical framework, robots are inherently *ethical* (and social), even before raising moral questions on their use or capacities, which must be distinguished from being inherent and being projected upon.

By outlining specific relations or disconnections in the literature, I start an ethical transformation through the very discussion I undertake – one that ideally concludes with a better understanding on emerging digital technology, such as robots, with the ability to reflect on their use and agency holistically.

Methodology

I undertake a comparative and critical analysis of ethical discourses around social robots, robots as humanoid companions, robotic machines, and digital algorithms by surveying Robot Ethics, Media and Surveillance Studies, Humanoid Robotics, and Machine and Computer Ethics (as aligned to wider Philosophy of Technology, POT).

First, I align ethical data concerns from MSS to morality-led Robot Ethics views, so as to reflect upon insufficient robot agency models in Robot Ethics and POT. Then, I highlight problematic tendencies in POT, on how morality is conceptualised into machines and algorithms, compromising moral agency in its relation to agent accountability. I outline the advantages and limitations in contemporary ethical streams to establish a new ethical discussion on robots, while keeping their companion position and their tracking agency as conflicting but equally important views.

My methodology makes use of an intersection between themes, layers, and frameworks. One layer moves from social robots towards algorithmic systems, as it traverses from the superficial, perceptive view on social robots into a computational, structural view on robots as informational systems – or from the robot body into the robot system. The second layer develops horizontally and reciprocally from morality to ethics agency discussions, to untangle conflations between ethical and moral perspectives on robots and their aligned agency models. These layers overlap throughout the thesis as I move to different levels of technological structure or ethical canon.

The themes are, for instance: the companion position of social robots; the robot's position as a data tracking device; anthropomorphism in HR; the single/appearance driven moral agenthood model; instrumental agency models of technology;

Posthumanist ethical views on technology; distributed agency models; tracking as an ethical process; the morality and ethics conflation; algorithmic agency networks; the advantages and disadvantages of distributed agency or computational interaction in robots; and the new challenges of human accountability.

These specific themes are used to explain exemplary frictions in the frameworks of (i) Robot Ethics, (ii) Media and Surveillance Studies (MSS) (both dominating in Part One), and (iii) Machine and Computer Ethics or POT (dominating in Part Two). These frameworks are the cornerstones of the investigation on agency models and ethics, but they incorporate further disciplinary views, which will offer more contextualisation of certain themes, as achieved through Humanoid Robotics and Posthumanism.

The frameworks reflect on this distinction between moral and ethical concerns that I mentioned as a challenge in the introduction. Robot Ethics (i) and Machine and Computer Ethics (iii) are morality-led discussions (as first and third framework), which means that while (i) is anthropomorphic, (iii) is anthropocentric. MSS (ii) (the second framework) allows for an ethical discussion to evolve; one that is supported by Posthumanist ethics and film-ethical concepts.

Within the framework of (i) Robot Ethics, I survey the ethical discussion on the use of social robots in elderly care as my case study. I claim that the morality-led focus in (i) Robot Ethics does not allow for ethical concerns of robots as a digital tracking technology to emerge, since this framework is highly anthropomorphic in its expectations towards robots. By surveying the literature in this segment, I aim to identify the anthropomorphic fallacy and to explore its dominance in the wider epistemological framework that Robot Ethics borrows from HR. This view favours

discussions on moral agency of robots as bound to their perceptive sociability, their humanoid design, and the expectations to act as companions rather than being only machines.

Within the framework of (ii) Media and Surveillance Studies, I argue for an additional ethical discussion on the computational capacities and concerns, as I focus on the ethical importance of data outlined through (ii) Media and Surveillance Studies (MSS). The perspective from (ii) Media and Surveillance Studies (MSS) proposes an exclusively ethical view that does not specifically look at social robots, but at one of their most concerning capacities: to collect and process data. Aligned to this framework, I establish an ethical and Posthumanist view on the data infringement/collection issues attached to the use of social robots as digital devices. Such a view would suggest the social robot as a network of non-evaluating, non-accountable agencies, as Posthumanists might consider. However, this perspective will challenge the question on robots being accountable for anything they do, elaborating on a theoretical complexity around the relations involved to form and use this technology.

The relevance of this investigation has severe practical implications for elderly care, as I briefly mentioned in the introduction. By linking (i) and (ii), I amplify the practical concerns robots present for elderly care as potentially immoral companions, but also as computer systems without immoral intentions, though with unethical capacities. As social robots primarily interact with the elderly, they not only manage data, but further collect it - with such data collection not critically reflected upon in Robot Ethics. What I am concerned with in this investigation is not only a theoretical neglect in Robot Ethics, but how newly formed issues around data infringement and health data ownership are overlooked; especially as I consider a non-consensual and opaque gathering of data for

potential commercialisation of such (Knoppers & Thorogood, 2017; Stahl & Coeckelbergh, 2016) to be fundamentally unethical. The introduction of more digital technology, in the amount, complexity, and in the form of robots into elderly care, has additional implications for the care profession, which are not ethically and fully explored.

Within the framework of (iii) Machine and Computer Ethics, as part of wider POT, I increasingly leave the focal point on social robots as humanoid bodies, as I theoretically move into their algorithmic agencies. This will still be a discussion on social robots, but it transforms into one that looks at these devices and their agency model differently: as a network of computational agencies. The difference to Robot Ethics is that these morality discussions are anthropocentric and do not argue for any appearance-based agency of robots. Instead, they look at the technological capacities of robots as machines and algorithms.

I explore themes such as artefactual and metaphorical agency, autonomy, accountability, moral decisions and computation decisions, single agenthood, distributed agency, and so on, to show how the second layer discussions between moral questions and ethical structures in POT seem to entangle and conflate, but differently to previous discussions. I investigate how newly conflated moral-ethical discussions become devoted to a Posthumanist ethical stream, but are not fully committed to questions on moral accountability and might contest both morality as a practical concept and its use for applied ethics.

The relevance of this investigation expands on theory in POT and, in a wider sense, outlines a thematic shift from applied ethics to meta-ethics. As I highlight how morality

research fuses with algorithmic modelling, I reflect on advantages and disadvantages of these streams. I examine how single agency models become increasingly disconnected from practical questions on digital systems, such as applied concerns on the use of robots, as these also become increasingly autonomous and complex in their use. What I assume is that, as more operational decisions are outsourced into the robotic or computational systems (away from the developer's decision), less concern is placed on asking simple questions, such as: why do we use robots? Who profits from their application? And who/what should be accountable for any potential undesired (to be defined) consequences? This part is concerned with a deeper understanding of how accountability and autonomy develop in the context of digital systems, and what new challenges await in this framework.

Chapter Structure

Chapter One. On Morality and Ethics – What Are the Differences?

In Chapter One, I offer insights from traditional moral philosophies/ethics (Blackburn, 2001; Ward, 2015) on why morality is an evaluative or reasoning concept. I then continue with two exclusively ethical perspectives on understanding technological agencies. This functions as the theoretical base for chapters Two and Three to further illustrate the concerns. I argue, at this point only, for a difference between moral and ethical questions. What I point out is that questions on moral agency in technology are bound to either an anthropomorphic projection or an anthropocentric evaluation of technology, so as to appear or act morally bound to (universal) human values, but that this hinders a wider ethical understanding of technology as a distribution of agencies - as Posthumanist theories suggest (Barad, 2003, 2007; Braidotti, 2006).

I also highlight that an ethical view of technology does not raise questions on practical accountability, which I consider essential for the discussion on social robots. However, since I believe that the network and intersection of technological and human agencies in robots are not fully understood in Robot Ethics, I include Agamben's (2000) work on early cinema as an ethical gesture to allow for a non-moral perspective on technological agencies to emerge. This is revisited in the section on tracking in Chapter Six.

Chapter Two: On the Anthropomorphic Fallacy and the Humanoid Companion in Elderly Care Seen Through (i) Robot Ethics

Chapter Two surveys (i) Robot Ethics to reflect on the companionship position of social robots in elderly care and on its ethical concerns and limitations. I examine this

framework to find out how robots are perceived and positioned by Robot Ethics and how the field understands what agent role they are given, and the purpose of them. In this chapter, I offer a cumulative summary of the literature in Robot Ethics (Royakkers & van Est, 2016; Sharkey & Sharkey, 2010; Turkle, 2005, 2007, 2011).

I amplify that Robot Ethics discusses social robots predominantly from an anthropomorphic angle, which means it derives concerns from their alikeness to the human body, not from their ability to judge or understand morality. I therefore consider their disciplinary ethical critique to be limited, because of its single anthropomorphic focus and because of its neglect of data-related issues that the robot creates by being a tracking device. Social robots are contextualised in this literature as problematic due to their companion position that allows them to deceive, harm, or spy on the elderly; what I consider to be a mostly anthropomorphic understanding of agency.

While the ability of social robots to track and detect human movements, emotions, gaze, or gestures does not affect the ethical discussion of data-related infringements, unless applied to a privacy-related intrusion, the implication of Robot Ethics is that the social robot is morally problematic – the robot can monitor, survey, or intrude as an agent.

Chapter Three: On Social Robots as Potential Dataveillance[9] as Seen Through (ii) Media and Surveillance Studies

I suggest in Chapter Three to look at the social robot as a digital device and align it to discussions on dataveillance from (ii) Media and Surveillance Studies (MSS). I justify this angle by viewing social robots as dataveillance, which I suggest discussing

[9] Dataveillance is the 'systematic use of personal data systems in the monitoring or investigation of the actions or communications of one or more persons' (Clarke, 1988). See Key Terms.

ethically through critical streams on digital technologies from MSS, reaching beyond the moral concerns raised in Robot Ethics (and their companionship view on humanoid robots). This chapter makes use of research by Gitelman (2013), Andrejevic (2012), Ball et al. (2012), and others, and amplifies why data-related privacy issues do not require a morally capable technology or an immoral agent, but are instead embedded in the making and emerging of tracking data in the first place, or in using robots for elderly care. This chapter is less concerned with social robots or elderly care, but with the data-gathering process that affects the abilities of social robots as digital technologies.

Chapter Three outlines that the collection of data is already an ethical and intentional process that affects Robot Ethics as much as any other digital, interactive device. Through this discussion, I show that ethical problems are embedded in the very collection of data, in designing tracking modules, and in the application of robots into elderly care. Finally, I show that these factors further create new problems for media literacy and data privacy in general.

Chapter Four: Epistemological Context on Morality, Agenthood, and Agencies of Technology

Chapter Four is a contextualising chapter in which I present wider techno-philosophical streams to offer insights into technological agency, human agenthood, and moral accountability.

First, I make use of a brief media-theoretical angle joined by a Posthumanist input to introduce the chapter, and then move to the main techno-philosophical perspective of Johnson & Noorman (2014). What I highlight through their work illustrates the concerns I see in newer autonomy and accountability debates, which increasingly

complicate technological and robotic agency as much as morality research. Their work allows me to reflect on technological agency as a metaphoric and instrumental concept, which I disagree with but still utilise to raise a new critique on the position of social robots as companions. In Chapter Seven, Wallach & Allen (2009), Arkin (2009), Brey (2014), and Floridi (2014) will push this debate further by pointing to the conflation and operationalisation of autonomy, agency, and accountability.

Chapter Five: Epistemological Context on the Apparent and Projection-Based Moral Agenthood of Social Robots

Chapter Five is a contextualising chapter on the limitations of the anthropomorphic fallacy that drives the mistaking of robots as moral agents. In this chapter, I show why the anthropomorphic view on robots is not only limited, but *limiting* within its own POT canons. I refer to an unreflected appropriation of moral philosophies into POT and examine the superficial and appearance-based view on moral agency around social robots through existing critical research by Coeckelbergh (2010). His work offers a techno-philosophical perspective on the superficial moral projections by bringing up the 'psychopathic' robot, a concept he develops on morally-appearing, yet unaware, robots. Through unpacking Kahn's et al. (2012) experiment on moral appearance, I offer additional and concerning proof on why building moral agency in robots based on appearance is problematic and why the anthropomorphic companionship position does not offer a fundament for moral values or actions to be derived from. With this chapter, I deconstruct companionship view on humanoid and social robots further through presenting the wider critique within POT streams.

Chapter Six: From Humanoid Robotics and Anthropomorphism to Computational Interaction and Tracking.

In Chapter Six, I situate the companionship and anthropomorphic fallacy by leaving the ethical and morality discussions on robots, and instead trace its origins in Humanoid Robotics (HR). At first, I consider the influence of HR on Robot Ethics a crucial hindrance for ethical discussions to emerge. This background indicates the (1) Moral Appearance and Agenthood limitation (which leads to robots being mistaken as humanoid companions, to be judged according to moral guidelines) emerges from the social paradigm that HR has established around social robots as companions and partners, aiming to leave their machinic side out of the discussion.

What I present is how social robots embody a threshold between the desire for human-like sociability and their status as sophisticated non-human machines. I do this by surveying the literature in HR, referring to authors such as Breazeal (2002, 2003), Fong et al. (2002), Lee et al. (2005), and Duffy (2003). I also explain that HR does not only discuss the anthropomorphic dimension of social robots as companions, but is also concerned with their computational sociability.

To allow for a transition between Chapters Six and Seven, I change from discussing the perceived sociability of social robots as humanoids to reviewing their computational abilities, such as tracking. This integrated shift, from anthropomorphic interaction to computational interaction, is not only mirrored in the shift between Chapters Two and Three, but is utilised to establish an ethical view on tracking aligned with Agamben's (2000) reflection on early cinema.

Chapter Six concludes with two additional side discussions. The first one presents a new and interesting anthropomorphic ethics on robots, named 'Synthetic Ethics' and suggested by Damiano & Dumouchel (2018). The second one is on a wider but very influential issue, which is around the rhetorical blurriness and imprecision in HR literature; this might further complicate agency discussions through rhetorical inconsistencies.

Chapter Seven: From Moral Robot Agents Towards Ethical Algorithmic Structures in (iii) Machine and Computer Ethics/POT

In Chapter Seven, I enter the discussions on machine morality research, as I demonstrate how POT streams steadily suggest new forms towards an algorithmisation of morality (moral decisions, actions, agency, etc.) at the cost of removing reflective and practical elements bound to applied ethics. I illustrate the tendency of machine or computer morality research to join and merge with the flattened, Posthumanist ethical streams that support a synthesis of the moral and the ethical discussions at first, but consequently seem to come with different issues. This chapter shows also that even if some ethical streams happen to relate to each other, these cannot be simply synchronised. Equally, the ethical disconnection between Chapter Two and Three amplifies a fundamental division in how technology, or robots, are understood in different areas; be it as a humanoid agency, moral entity or as an ethical network. Therefore, this chapter does not synchronise the previously pointed out, but leads to new consequences that harm the morality discussion, albeit differently than I expected.

In Chapter Seven, I point to advantages and disadvantages in understanding how moral agenthood is translated into robotic or computational systems; then, I highlight two

more limitations when conceptualising moral agency within technology, which all favour a techno-operational structure but not a reflexive, or, as Ward (2015) suggested, 'dynamic' (7) one.

The first, on a superficial apparent morality model, was mentioned in Chapter Five. The next two are not discussed in literature, yet these emerge from my own critical reading. I identify a second limitation through parameters supplied by Wallach & Allen's (2009) work that favours the conflation of moral autonomy and computational autonomy leading to a (2) Reductionist Morality. Then, by thinking (3) Distributed Morality, I use the work of Brey (2014) on 'structural ethics' and Floridi's (2014) work on 'mindless' morality to illustrate to what a distributed and flattered view on moral agencies into algorithms. Even if I speak of limitations in thinking morality and moral agency, I acknowledge as much that each of these discussions leads to both positive and problematic developments. The conclusive concerns I raise from this chapter revolve around accountability not being debated critically enough in POT. This position seems to be left unresolved in the context of moral agency, which affects wider applied ethics streams and, ultimately, influences practical and legal questions around robots.

Contribution

I establish a new and concerning linkage between social robots in elderly care and their potential as dataveillance by connecting Robot Ethics with Media and Surveillance Studies, so as to present a holistic and ethical reflection on the practical consequences around data collection and privacy infringement in the use of robots. Through linking Posthumanism with Philosophy of Technology, I further contribute to the ethical and moral research around robots and algorithms by challenging: the anthropomorphic and anthropocentric tendencies to discuss robot and algorithmic agencies; the reductionist appropriation of morality to code; and the increasing inability of techno-ethical streams to address accountability concerns when placing robots into real-life environments.

In Part One, I contribute to Robot Ethics and Philosophy of Technology by forming a connection between dataveillance discussions in Media and Surveillance Studies and morality-led discussion on social robots in Robot Ethics. I contribute to two frameworks: (i) Robot Ethics, where I draw attention to data-related ethical issues situated in the use of social robots, and (ii) MSS, where I encourage the inclusion of social robots as data tracking devices in their critical canon. Linking these frameworks is crucial to address the practical consequences of social robots for elderly care and elderly patients on a new level of privacy and data infringement, which are not sufficiently addressed in existing morality-led streams.

Especially, the growing concerns around the commercialisation of health data and the ubiquitous applications of digital technology into sensitive environments highlights the value of this thesis for various disciplines dealing with tracking and privacy, or for

contexts such as professional health care industry concern of how and if to apply robots in care.

In Part Two, I contribute to a wider ethical discussion in Philosophy of Technology by formulating new concerns in the alignment of moral agency to digital systems. By pointing to the limiting anthropocentric discussions in the (iii) Machine and Computer Ethics and POT framework, I highlight conflations of operational and moral autonomy, perceived and inherent agency, and of functional responsibility and moral accountability in algorithmic systems.

First, I detect new dangers in aiming for a reductionist and instrumentalist conceptualisation of morality as applied to algorithmic rules. Second, I amplify how moral reasoning and norming is increasingly detaching itself from accountability questions, which I see as a side effect of the alignment of morality to code. This development not only implies a problematic return to a techno-moral positivism, which might impact models on human morality, but also allows to conflate operational autonomy in digital technology with moral autonomy; the latter is always lacking in robots and algorithms.

I identify that the avoidance to address uncomfortable questions in POT - on who accounts morally, ethically, and legally for robots or digital systems - shows in the dissolving of essential components when thinking of moral agency as a dynamic and critical discourse. I stress that this will have practical consequences for various fields, as much for elderly care as for environments in which robots or tracking systems are being trusted by their users.

I. ON MORAL AND ETHICAL CONCERNS AROUND SOCIAL ROBOTS

Part One summary: Part One surveys wider ethical and moral conversations on the potential misuse or misbehaviour of social robots as moral agents or as ethical structures in elderly care. By distinguishing between a morality-led and ethically-exclusive way to think of technology, I provide the reader with an initial understanding that these terms can lead to different epistemologies in comprehending technology. I will highlight that thinking of social robots through either a morally-led or ethically-led focus has advantages and disadvantages. After introducing what morality- and ethics-led discussions can look like, I will survey existing ethical discussions on social robots. First, I survey the ethical literature on humanoid companions in elderly care and (i) Robot Ethics. Second, I discuss their position as tracking devices from a (ii) MSS perspective on data gathering as ethically concerning. What I amplify is that a neglect of the ethical dimension of social robots as tracking machines, as witnessed in Robot Ethics, is problematic when it comes to data, privacy, and literacy of the elderly, but it also builds on the dominance of treating social robots like companions.

1. From Thinking Morality to Posthumanist Ethics

Chapter summary: Chapter One reflects on the inadequate distinction between morality and ethics in traditional moral philosophy and as it presents new Posthumanist angles to think of technology ethically (Barad, 2007; Braidotti, 2006; Agamben, 2000). I hereby introduce my concerns with thinking technology through an exclusively moral lens. I make use of the differentiation between ethics as a structuring concept, and morality as an evaluative concept, to highlight that morality discussions on technology are problematically projecting concerns from questions around human agency into questions on technological agency. As I lay out critical views on moral reasoning (Luhmann, 1989) and the institutionalisation of morality (Laidlow, 2014), I also highlight the capacity of traditional moral philosophy to practically assign moral accountability; something that the ethical canons of Posthumanist ethics are not able to do in practical terms.

I would like to introduce certain distinctions between morality and ethics to amplify deficits in the discussion on social robots, which would otherwise not crystallise as such. I critique that the moralisation of technology allows for two things to happen: firstly, the anthropomorphising of moral qualities; and secondly, the anthropocentric expectations of computers or robots to decide or judge morally. My critique of anthropomorphism dominates Part One, and while anthropocentrism encapsulates Part Two.

I differentiate in this thesis between ethical and moral agency, but this kind of delineation is barely undertaken in Philosophy of Technology or in Robot Ethics (Royakkers & van Est, 2016; Kroes & Verbeek, 2014). Mostly, ethics and morality are conflated. In principle, these are interchangeable terms in Philosophy (Kroes & Verbeek, 2014; Luppicini, 2009). By way of example, I want to mention Luppicini (2009), who writes that ethics 'is the study of moral conduct, i.e., conduct regarded as right and wrong, good or evil, or as what ought or not ought to be done' (19). However, for a Posthumanist philosopher, ethics is not the study of morality at all. Theoretically, the conflation of ethical and moral questions is not uncommon in philosophical streams per se (exceptions include Braidotti, Barad, or Introna), but it can become one in practical terms. As I will emphasise, a morality-led debate can take a different angle than one that is ethically-led. It makes sense to further distinguish between meta-ethics and applied ethics (Luppinci, 2009; Ward, 2015), since these have different goals.

I want to first introduce wider philosophical differentiations between morality and ethics. As aforementioned, morality research is often referred to as *ethics* (see Robot Ethics), while the research can still be only morality-led, since these terms come together. Again, morality discussions are always ethical, but there are ethics that refuse

morality as a concept or model. What this shows is that *ethics* is the structural vessel in which moral relations, virtues, actions, or norms of individuals or societies (or technology, as in this thesis) are debated, but debating evaluations or projections of good or bad behaviour is not always relevant to understanding how ethical relations unfold. Hence, what makes this topic slightly confusing is that even when philosophers discuss ethical structures, they can still mean and reflect on moral questions; however, in technology, this conflation can be problematic and less useful ethically.

Ward (2015), a media philosopher, approaches ethics from what he calls a *radical* media philosophy, by offering a semi-distinction on ethics and highlighting the evaluative process of ethical activities (what I perceive to be him talking about moral evaluation). He writes that:

> 'Ethics is the study and practice of what constitutes the best regulation of human conduct, individually and socially. Humans apply their notions of ethics by acting according to principles, norms, and aims. Ethics is the activity of constructing, critiquing, and enforcing norms, principles, and aims to guide individual and social conduct. The phrase "the best regulation" indicates a zone of critical and ever-evolving thought about the notions and norms of ethics. Existing norms may be inadequate, or even unethical' (4).

I agree with what Ward highlights, that ethics is a *practice* of regulation leading to norms, which can be potentially unethical/immoral. However, the practical side of ethics is important for him as well; ethics is a way to provide society with tools on how to live as a community and how to conduct oneself in society. Ward also conflates ethical and moral intentions, but what is important to him is also what I will keep highlighting; both terms are active and dynamic processes, not fixed schemata or strict rules. Hence, for Ward (2015): 'Ethics is practical. Ethics is an activity, a process, and a dynamic practice. It is something we do. We do ethics when we weigh values to make a

decision' (5). However, further, he states that 'ethics at its best is reflective engagement' (6). Ward's definition offers a solid but not distinct orientation between ethics and morality, and I will return to his views multiple times to support my critical view upon the reductionist view on moral decisions around the algorithmisation of morality (see Key Terms section).

Blackburn, on the other hand, clearly formulates differences between a moral and an ethical climate. In his book, *A Little Introduction to Ethics* (2001), he claims:

> 'An *ethical* climate is a different thing from a *moralistic* one. Indeed, one of the marks of an ethical climate may be hostility to moralizing, which is somehow out of place or bad form. Thinking that will itself be something that affects the way we live our lives. So, for instance, one peculiarity of our present climate is that we care much more about our rights than about our 'good'. For previous thinkers about ethics, such as those who wrote the Upanishads, or Confucius, or Plato, or the founders of the Christian tradition, the central concern was the state of one's soul, meaning some personal state of justice or harmony' (3, 4).

What Blackburn points to, and Ward does equally, is the structural element in ethics and the moral evaluation as 'out of place or bad form' (4). I consider that these two *climates,* to which he refers, address different levels of discussion, but do not exclude each other. Yet, conflating these two might only lead to an unnecessary moralisation of technology. Accordingly, as will be later argued, computational systems have further complicated what moral norms refer to in the context of technological autonomy.

It is important to note that neither Blackburn (2001), nor I, imply that a differentiation is necessary because either morality or ethical questions are superior to one another. The terms are dependent on the discussions, which often treat them as interchangeable because it does not make a difference. My argument is not that holding on to anything related to morality has become obsolete in the context of technological agency.

However, I am concerned with how certain theories are instrumenting morality; I stress that the concept must be renewed and taken out of the instrumental and reductionist corner - otherwise, as is happening, Philosophy of Technology (POT) discussions are in danger of relapsing into already questioned and overcome dualistic and simplifying morality camps. What I encourage is a view beyond moral agency or reasoning, to recognise that there is an *amoral* but still ethical dimension of social robots; one that can be accessed only through sidestepping the moralisation of technological intention or autonomy; one that shows there are inherent ethical structures and consequences in place prior to human immoral intentions. One reason for distinguishing ethics from morality is that the evaluation process Blackburn talks about asks for reasons and for accountable agents.

However, morality presents other concerns even before being aligned to technology. According to the sociologist Luhmann (1989), it can be problematic to hold on to morality as an evaluating system, because we cannot agree on why humans *should* be moral. He reviews the concept of morality as 'deficiency-oriented'. For him, the search for moral reasoning or justification remains an indicative search for something inherently good or bad in human agents. In Luhmann's view, whenever the catchword *morality* appears, the experiences Europe has had with morality since the Middle Ages emerge as examples; religiously adorned upheavals and suppressions; the horrors of inquisition; and wars about morally-binding truths and revolts arising in indignation (1989: 370).

Luhmann (1989) views ethics and morality from a sociological and systematic perspective, defining ethics as a 'reflective theory of morality' (360); as a structure in which morality develops, which relates to Blackburn. However, Luhmann does not

intend to provide guidelines for a practical morality or to be useful in exploring robots ethically. Ethics is, for him, a reflective theory of morality, able to address a responsible and accountable interaction beyond imperatives of good and bad. The enquiry towards answering and understanding moral reasoning is, for Luhmann, an:

> '"[E]thical" task – and ethics for him means: reflection on morality (as a phenomenon of communication). Luhmann's view on ethics is not about the formulation of some basic moral rules or "imperatives", or judging morality as morally "good" or "bad"' (Moeller, 2006: 112).

Most importantly, Luhmann concludes that the ethical debate around finding acumen for moral reasoning has never produced any answers and no actual reasons for morality in the first place. 'Ethics can't provide reasons for morality. It finds morality to be there, and then it is confronted with the problems that result from this finding' (1989: 360). Luhmann's grasp of the error of looking for a moral reason for an agent helps to explain why moral behaviour/reasons and ethical structures are not the same discussions.

This comes close to the aim of my thesis, which is on the constitution of moral norms and ethics within techno-ethical streams, by also offering a critique on a superficial or projective moralisation of technology. I see Luhmann's resistance towards morality and moral reasoning in his denial of universal and institutionalised moral norms; hence, I agree with this distrust. I also acknowledge that Luhmann, in being a Structuralist, is highly concerned about the relations and codes in which systems operate, so his view on human relationship is not interested in evaluations, but in structures.

Similarly oriented towards understanding ethical structures, Foucault differentiates between ethics and morality (cited in Laidlow, 2014) to amplify the institutionalisation of morality as an important aspect. For Foucault, the moral codes are, as much as for Eshleman, not simply individual judgements that are made by a human deciding for

themselves on what is *good* or *bad,* but institutionalised rules that mesh with the accountability of the agent. Laidlow (2014) writes:

> 'Foucault distinguishes between what he calls moral codes—rules and regulations enforced by institutions such as schools, temples, families and so on, and which individuals might variously obey or resist—and ethics, which consists of the ways individuals might take themselves as the object of reflective action, adopting voluntary practices to shape and transform themselves in various ways [...]. Ethics, including these techniques of the self and projects of self-formation, are diagnostic of the moral domain' (29).

Therefore, for Foucault, the ethical is bound to encompass 'reflection' and 'self-formation', and can aid in understanding ethics as processuality or unfolding, and not as a stable regulation of codes, even if these might result as a consequence of this process. According to Foucault, ethics reflects on the moral codes by being the process of their negotiation. However, the difference would be in saying that the moral sphere is considered the institutionalising of norms, which might be physically installed into tracking as values and descriptions of behaviour. However, these values would not mirror the structure that, for instance, tracking, not technology in general, is establishing and unfolding.

As pointed out by Levinas & Hand (1989), Levinas also does not support a single and stable agenthood ethics, but understands ethics as, in itself, a relationship that unfolds between two people. He emphasised that the entanglement between 'I' and 'Other' is an already *ethical* process[10] without yet having to bring in the evaluation or reasoning. This relation is inherent in being human and the 'primacy of relation explains why it is that

[10] 'Ethics arises from the presence of infinity within the human situation, which from the beginning summons and puts me into question in a manner that recalls Descartes's remark in his third Meditation that 'in some way I have in me the notion of the infinite earlier than the infinite.' Consequently, to be oneself is to be for the other. Levinas has summarized this fundamental point in an article entitled 'Beyond Intentionality'' (Hand in Levinas & Hand, 1989: 5).

human beings are interested in the questions of ethics at all. But for that reason, Levinas has made interpretative choices' (Bergo, 2011). An example by Van de Poel & Royakkers (2006) illustrates this as a difference between applied ethics and ethical theories, arguing why ethics needs to think beyond moral decision-making. They argue:

> '[A]pplied ethics is essentially the application of general moral principles or theories to particular situations (…). Different theories might yield different judgments about a particular case. But even if there would be one generally accepted theory, framework or set of principles, it is doubtful whether they can be straightforwardly applied to a particular case. Take a principle such as fairness. In many concrete situations, it is not clear what fairness exactly amounts to' (2).

Kroes & Verbeek (2014) point to why we even consider technology or artefacts to embed moral agency in the first place. This idea can be traced back to the Enlightenment, when the theological views on morality were transferred from God to humans, which distributed the moral responsibility to the individual and away from the superior power. After the Enlightenment, God became less accountable for human misfortune. Yet, a new fear surfaced at the same time (Dumouchel & Damiano, 2017) that robots could form harmful intentions, even if this is the most unlikely scenario for now. Therefore, humans have pushed their acquired empowerment one step further, by transferring moral agency to material agencies, which they have built in a *God-like* manner. Kroes & Verbeek (2014) state:

> 'This 'material turn' in ethics raises many questions, though. Is the conclusion that material things influence human actions reason enough to actually attribute morality to materiality? Can material things be considered moral agents, and if so, to what extent? And to what extent can artificial moral agents be constructed with the help of information technology? The attribution of some form of moral agency to technical artefacts not only requires a rethinking of the notion of agency but also of morality' (4).

This ambition might still be a hypothetical one, but it attracts lots of research and occupies various ethical streams in HR, POT, and Robot Ethics equally. From this attempt, new concerns emerge on how to think of morality in material and non-human terms. Major issues I see in morality-led conversations on technology are that these often favour an anthropomorphic and anthropocentric expectation of the human developer, designer, or programmer, but end up supporting universal or simplified views on what is the *good* or *bad* in the very technology. I will repeat these concerns a few times, since they occupy my whole investigation. If agreeing that morality is understood as a concept that evaluates *good* and *bad* action or behaviour, implying that something is moral or immoral suggests two things. Firstly, it implies an evaluative process and, secondly, it implies an agent taking responsibility for the behaviour.

Therefore, what might be an important aspect in holding onto morality – in the context of technology – is to define new standards of agencies and accountabilities, as I would conclude. In principle, moral intention or action requires the allocation of an accountable agent, and, for most cases, this is a human agent (Eshlemann, 2016; Brey, 2014). Considering the institutionalisation of morality is bound to agent accountability (as in religious or legal systems), only the (human) agent is held accountable for his or her actions or the consequences of such. Thus, immorality is not something that rests unnoticed in these systems; immorality has consequences, which are punishable (Brey, 2014).

If thinking about the Ten Commandments, these might be guidelines on how to behave correctly, but not only for the sake of being a nice person, but for sake of living in a peaceful society and in the context of a religious system, which knows sin and guilt to punish the disobedient agent with.

Brey (2014) considers the concept of moral responsibility as linked to accountability and agency. For him, moral behaviour and actions are related to and depend upon the ability to be blamed, punished, or rewarded for these – in short, to be held accountable. For him, 'moral responsibility from our conception of moral agency is also unappealing' (2014: 134). While Brey (2014) highlights the links between moral agency and moral responsibility, Eshleman (2014) points out the importance of the responsibility of the moral agent by writing:

> 'A person who is a morally responsible agent is not merely a person who is able to do moral right or wrong. Beyond this, she is accountable for her morally significant conduct. Hence, she is an apt target of moral praise or blame, as well as reward or punishment'.

Thus, the questions on agenthood are closely aligned to these on moral accountability. However, the more that technological systems are understood as distributed systems, which question any agenthood position (as Posthumanist ethics but also the discussions in Chapter Seven will show), the more complicated Eshleman's remark on accountability, blame, and punishment will be in practical terms. It becomes harder to discuss how morality cannot be assigned to any agent.

Responsibility and accountability are hugely important drivers behind traditional morality models and cannot be simply removed as one major element in morality discussions. The issue when thinking that moral norms are rules is that these might seem perfectly transferable to computational codes, as Chapter Seven will show, but the robot, machine, or algorithm will nonetheless lack the relevant consciousness or awareness to understand punishment, the consequences of not executing these, and of immoral consequences. POT discussions challenge this agent position, but continue to be Universalist, so the question is; who is rewarded or punished? The increasing

autonomy of robots and digital systems is yet unable to link accountability to autonomy *technically*. For now, I see an agreement in POT and Robot Ethics, that no technology can be held practically accountable for its actions. Yet, in the context of applied ethical canons on robots, this is an uncomfortable truth that is hard to deal with practically.

To conclude, I amplified that ethics often refers to a structural vessel of morality in which to be discussed, but, as I will show next, there is an *amoral* way to discuss ethics. Morality is not only an important concept for human moral philosophy; it is also a hugely critiqued one in ethical streams on non-human/human agency relationships. Some would argue that morality is such a difficult, inherently Universalist and institutionalising concept; it is not suited to fully understand human agency, and neither should it be aligned to technological agency (Braidotti, 2006). Since aligning norms and commands from traditional moral philosophy to human agency is already difficult, it is even more complex to decide on how to 'computerize Kant's categorical imperative', as Wallach (2010: 247) points out.

Next, I will bring up the ethical angle around technological agency, which is less evaluative as interested in the entanglement of human and non-human agency networks, allowing for a non-moralising view.

Posthumanist Ethics: A Better Way to Understand Technological Agencies?

Early on in this research, I expected to find much more Posthumanist views around robots in Robot Ethics, since this field deals with new shifting paradigms that affect and lead to new concepts of a human/non-human agency intersection. However, I could not

find any.[11] This was a surprise, since in humanities-led discussions around techno-material agencies, digital technology, and data, Posthumanist theories are encountered frequently. This is also true in fields such as Affect Studies (Massumi, 2002; Gregg & Seigworth, 2010), Performance (Philosophy) Studies (Ruprecht, 2017; Dimitrova, 2017), and Cultural Studies (Ahmed, 2004; Parisi, 2013; Blackman, 2012).[12] My expectation came from knowing that Posthumanist ethics – without really giving much importance to the social robot – also dispute material agencies of technology/human/environment as intertwined (Braidotti, 2006; Barad, 2003, 2007) and address ethical concerns, which do apply to robots and algorithms equally.

From my initial survey of Posthumanist camps, I was aware that Posthumanist philosophers deny moral thoughts around agency debates (human- or technology-related), since they disagree with the moralisation of technology or humans fundamentally. They critique anthropocentric views on human or non-human structures and universal moral guidelines, expectations, or projections, which would not say much about the actual capacities of an artefact, technology, or robot (Braidotti, 2006; Zylinska, 2014). Instead, in the case of robots or digital technology, Posthumanist camps would rather suggest looking at the complexities of a computational distribution of agencies, without thinking in clear hierarchies or charging agencies with moral expectations, nor with aligning these to human agency.

The work of philosophers such as Rosi Braidotti and Karen Barad will be in the focus of my investigation on how to think of technology such as robots ethically. Braidotti

---

[11] I exclude AI discussion from its review.

[12] These disciplines have left their disciplinary grounds to join their research on common debates, such as affect and emotion, or gender and race. Therefore, researchers such as Blackman (2012), Ahmed (2004), or Barad (2003, 2007) are situated in multiple discourses.

(2006) and Barad (2003, 2007) present a non-anthropocentric angle on thinking technological agencies as one way to think about technology and human agency as equally important agencies in a non-evaluative way. However, there is one issue when following these theoretical camps; practical accountability is not debated to the extent that applied ethics would require. Posthumanists do not deal with the practical issues of why a social robot might or might not slap or push an elderly patient - not that Robot Ethics does sufficiently - but they are concerned with finding out what intention the robot might have or still lacks. Braidotti (2006), while not talking about robots per se, but about agents, is interested in shifting the agent position towards 'a nomadic subjectivity that involves a materialistic approach to affectivity and a non-essentialist brand of vitalism' (4). This relates to the metaphor discussion brought up with Johnson & Noorman (2014) in Chapter Five.[13]

When one summarises carefully and amalgamates these debates, a common thread becomes evident; morality is not a reference point in Posthumanist discussions. Nevertheless, this absence is not an oversight or an encouragement to argue that technology is immoral or that morality is superseded as a value system. This is because these debates lean towards a discussion on 'agential' relations (Barad 2003) and cannot find the theoretical purpose of moral agency as a single reference point.[14]

---

[13] Through Johnson & Noorman's (2014) work on technology being a metaphorical extension of the human, they consider that the humanoid design could be the *wrong* metaphor picked to define a social robot's agency. This implies that the level of *aliveness* in the robot might not be situated where its *humanness* is; at the perceptive level of the robot appearing to be human-like. The human visual alikeness might be the most *unhuman* part of the robot, in the end.

[14] Barad (2003) has coined the term 'agential realism', which refers to looking at the materialization 'of *all* bodies — "human" and "nonhuman"—and the material-discursive practices by which their differential constitutions are marked. This will require an understanding of the nature of the relationship between discursive practices and material phenomena, an accounting of "nonhuman" as well as "human" forms of agency, and an understanding of the precise causal nature of productive practices that takes account of the fullness of matter's implication in its ongoing historicity' (810).

Braidotti (2006) not only argues the limitations of a stable agenthood and morality, but also intentionality as a criterion in ethics. Braidotti's book, *Transpositions: On Nomadic Ethics* (2006), offers an in-depth discussion on the shift towards 'ethical complexities' in the debate of intertwined technological and human agencies. Braidotti suggests thinking in 'nomadic ethics' (15) to reject morality as a concept that 'highlights the relevance of a non-unitary vision of the subject' (15). She claims that such has gained 'importance of the notion of individualism in moral philosophy' (12), which is also critiqued by Zylinska (2014) as 'anthropocentric moralism (where values are being laid out without questioning the process of their fabrication and the conflict in which they always exist with some other values)' (72). Braidotti (2006) clarifies:

> 'The ethics of nomadic subjectivity rejects moral universalism and works towards a different idea of ethical accountability in the sense of a fundamental reconfiguration of our being in a world that is technologically and globally mediated' (15).

This must be done because 'moral philosophy is of hindrance, not of assistance, in dealing with the ethical complexities of our times' (15). This argument is crucial in this thesis, since it points to a provisional understanding of how the mediation within technology is an ethical process that forms accountability, in which the moral question could be a hindrance, not an endeavour. I consider it a huge problem, as I will point to in Chapter Two, that, on the contrary, Robot Ethics struggles to acknowledge the agency of the robot as a robot, but anthropomorphises its ethical issues from being a *perceived* agent (because of its humanoid shape) that is meant to act as a companion or friend to the elderly. On the other side, the robot is reduced to an instrument that can be misused by an immortal human agent. Both views, as I argue, are stuck in anthropocentric and moralising traps, as I will elaborate further in Chapter Two.

While Braidotti's views are theoretically valuable to understand the value of non-human agencies, and functions as my antithesis to the morality-focused research in Robot Ethics, it remains unclear how she debates practical accountability. The question on accountability is not simply overlooked; it is not addressed in the very practical case of moral accountability, because morality is not a factor any longer. Braidotti (2006) sums this up with:

> 'Ethics in [P]oststructuralist philosophy is not confined to the realm of rights, distributive justice, or the law, but it rather bears close links with the notion of political agency and the management of power and of power-relations. Issues of responsibility are dealt with in terms of alterity or the relationship to others. This implies accountability, situated-ness and cartographic accuracy' (12).

However, she is surprisingly clear that it is the 'awareness' of the subject who fills the accountable position after all, by also amplifying that such position is ideally a critical one, as well as stating that 'ethical accountability is closely related to the political awareness of one's positions and privileges' (13). However, her views (and that of other Posthumanist researchers) present their own limitations, in being unclear on practical discussions around the use of robots or any other technology.[15]

However, the newly attained theoretical complexity is worth highlighting, because this influences how to understand the practical consequences as much. Karen Barad's work (2007) introduces the neologism of 'ethico-onto-epistemology' (409) as an inherent entanglement of technology, human, and other material agencies - and as *ethical*. She writes that her terms:

---

[15] It is important to point that the Posthumanists are not necessarily indifferent to questions on responsibility or accountability questions (Braidotti, 2006). But they do not enter an applied ethics canon, neither do they join canons around practical issues in the use of robots or on accountability questions emerging from such. This does not mean there could not be a Posthumanist view on robots (as agency network) that allows to theoretically debate for a holistic accountability concept.

> '"[O]ntoepistemological" marks the inseparability of ontology and epistemology. I also use "ethico-onto-epistemology" to mark the inseparability of ontology, epistemology, and ethics. The analytic philosophical tradition takes these fields to be entirely separate, but this presupposition depends on specific ways of figuring the nature of being, knowing, and valuing' (409).

While acknowledging that these three methods and disciplines are always intertwined, the entanglement of human, material, or technological agencies is understood as being *performative*[16]. Barad refers hereby to Butler's concept on 'gender performativity' that puts the focus on the *making* of gender, not the *being* of gender. Applied to tracking, maybe this allows us to understand it as a *making* of ethics, not the *being* of moral codes.[17]

To conclude this section, morality is never a reference point in the post-anthropocentric, or Posthumanist discussions on technology (or any other agencies - human, material, or environmental - since these are understood as intertwined), because of the awareness of how morality is shaped by other value systems around it, and how it ends up being instrumented or institutionalised. An ethical perspective not only removes the evaluation process from understanding the intersection of relations, but, as Barad suggests, Posthumanists focus on the link between ethics, epistemology, and ontology when thinking of agency.

---

[16] Barad (2003) proposes 'a posthumanist notion of performativity—one that incorporates important material and discursive, social and scientific, human and nonhuman, and natural and cultural factors' (809). She relates this to the work of Butler on performativity and materialisation of gender.

[17] 'The notion of performativity has a distinguished career in philosophy that most of these multiple and various engagements acknowledge. (…) Butler elaborates Derrida's notion of performativity through Foucault's understanding of the productive effects of regulatory power in theorizing the notion of identity performatively. Butler introduces her notion of gender performativity in *Gender Trouble*, where she proposes that we understand gender not as a thing or a set of free-floating attributes, not as an essence—but rather as a "doing"' (Barad, 2003: 808).

What these streams highlight, and what I agree with fully, is that defining *good* or *bad* values is bound to the epistemes deciding, similar to when deciding on human agenthood or human subject as rational entities. My critique of Robot Ethics and POT streams, which I reflect on in Chapter Two and Three, will circle around the norming of morality and agency as implied evaluations and power structures, which are blindly reproduced or assumed to be universal. I view these aspects as strategic instrumentations of a valuable reference, such as morality, and removed from the interest to make morally justifiable robots.

I consider the ethical view as more appropriate to understanding that responsibility and accountability are inherent in the forming of relationships, and that such do not operate in a top-down manner, neither with cleared stated values, which can be assigned or extended into the technology. If we instead acknowledge the wider network of agencies and reflect on their making, this might allow for better insights into what technological agencies are, then, the common fixation in POT on making a robot *good* or *bad* or human-like. Yet, I also acknowledge that without clear answers and stable points on accountability and responsibility when using technologies, especially, robots will lack the necessary control as much as an ethical understanding.

The Posthumanist approach is not fully ignored in newer POT canons (Introna, for instance, works closely around Barad's work). Chapter Seven will return to this point with Brey's 'ethical structure' and Floridi's 'distributed moral agency'. However, there is no resonance on this ethical perspective in Robot Ethics, which might encourage further questions to emerge, for instance, on the influence of social robots for professional care, or its redefining of caring, company, or trust, or the effect it might have for the elderly and their self-worth, health, or sociability.

As I will discuss in Chapter Two, ethically seen, the use for social robots does lead to an implementation of a new data management structure into elderly care. I show in Chapter Six that the robot's ability to track can be understood ethically and not morally to receive a full picture on the social robot as an ethical process of sociability.[18]

On Agamben's 'Ethical Gesture' as an Amoral View on Technology

To further the discussion on how to understand tracking as an ethical process, I decided to include Giorgio Agamben's (2010) work on ethical gesture in this debate. I do this, not to start a discussion on early cinema, but to suggest an example from film-philosophy that thinks of the socio-material agencies of technology, on the grounds of them being ethical gestures. The specific reason for dedicating space for an ethical understanding of cinema through Agamben's work is that it aligns the discussion on computational architecture and tracking in an ethical way in Chapter Six. This angle, even if it leaves the major discussion slightly, offers an alternative, even if speculative, take on the dominant moral debates. However, it will encounter limitations within practical questions in the use of robots beyond the human agent as the only accountable, and yet, undiscussed one. Chapter Six will explore this ethical view by outlining the making of tracking modules through Hariatoglu et al.'s (2000) work, and by expanding on a highly influential interdisciplinary concept on emotion detection, the *Facial Action Coding System (FACS)* by Ekman (2003).

[18] Barad argues: 'Crucial to understanding the workings of power is an understanding of the nature of power in the fullness of its materiality. To restrict power's productivity to the limited domain of the "social," for example, or to figure matter as merely an end product rather than an active factor in further materializations, is to cheat matter out of the fullness of its capacity' (Barad, 2003: 810).

Agamben's work, less as a cinema expert but in a wider philosophical sense, supported my realisation that robots are not understood fully in their remediation capacities; in this sense, they are not different from cinema or other media technologies in my view. Again, I have one clear intention with this example: to offer an ethical view on how to understand the making and entangling of agencies outside an anthropocentric or anthropomorphic focus and to point out that moral questions are not inherent to comprehending technological agencies, and still, might be useful to understand their use and application in real-life scenarios.

Agamben's (2000, 2014) conceptualisation of early cinema[19] as an ethical gesture allows for an understanding of technology as relational negotiations of agencies, and as a gestural remediation of socio-cultural and technological discourses. Understanding the emergence, development, and technology of early cinema as *gestural* and *ethical* has influenced multiple debates in film-historical research (Gronstad & Gustafsson, 2014), but it has also influenced Performance Studies and Performance Philosophy (Ruprecht, 2017; Dimitrova, 2017).[20] Interestingly, Agamben's work has been selectively anthologised in edited collections dedicated to visual and media culture, as much as it has appeared in a Performance Studies context (Harbord, 2016: 8), but he is not aiming for a media- or technology-centred perspective on ethics.[21] Instead, he is interested in a politico-ethical view on cinema by looking at how cinema technology has embodied

---

[19] Early cinema is used in this context to pinpoint a time frame during the early 20th century as important when technology, cultural, and other practices come together to form cinema as a technology. This *moment* is not understood as a singular one, and cinema is not seen as *invented* in one moment or as a concluded technology according to Punt (2000). Agamben is fluid between referring to cinema as a *dispositif* or as a technological apparatus.

[20] Performance Studies influenced this exploration as much by enabling a rethinking of dichotomies of appearance and qualities, agents and tools, bodily expression and virtues, and of ethics and morality.

[21] Agamben's engagement with media and film theory can be traced in his text, *Releasing the Image: From Literature to New Media* (2011).

and transformed human gestures, within a philosophically and historically complex revisiting of what he calls the 'crises of representation' in the early 20th century.[22]

In *Notes on Gesture* (2000), Agamben offers two crucial points on the contextualisation of early cinema as an ethical gesture. First, he states that the (bourgeois) gesture has been lost at the end of the 19th century, and, second, that early cinema restored it by *becoming* a gesture itself. For him, bourgeois gestures were based on the illusion of subjective identity and unity, on fixed and individual agents (ironically, a fallacy the social robot is trapped in as well, according to Dumouchel & Damiano, 2017).

Agamben argues that the *mediality* of (early) cinema had positioned cinema as an *ethical*, not aesthetic or representational, practice.[23] The ethical is, for Agamben, within the emergence of early cinema technology that unfolds as mediality and as a gesture of the discourses it incorporates - without being moralising in itself, but by being a biopolitical entanglement. His work contributes to this discussion by unravelling ethics as an intertwined negotiation of technological capacities, human aesthetics, and values. Agamben's exploration reaches beyond moral outcomes or human intentions in the unfolding of socio-technological practices such as cinema, but he sees this loss of

---

[22] The media philosopher Flusser opens a different discussion on why and how human gestures and technology are intertwined ethically (Flusser & Roth, 2014). Flusser looks at human gestures in their phenomenological expressivity (but also as a process, not as an expression), pointing to the difficultly in tracing the meaning and intention in causality (3). For him, technological media practices such as photography (Flusser, 2000: 33-41), have themselves created new human gestures, fusing the technological functions of the apparatus and the human gesture of using it.

[23] 'Gilles Deleuze has argued that cinema erases the fallacious psychological distinction between image as psychic reality and movement as physical reality. 'Every image, in fact, is animated by an antinomic polarity: on the one hand, images are the reification and obliteration of a gesture (it is the *imago* as death mask or as symbol); on the other hand, they preserve the *dynamis* intact (as in Muybridge's snapshots or in any sports photograph)' (Agamben, 2000: 54).

gesture in parallel to the decline of the bourgeois class and the gesticulations present in 19th century mannerisms (Ruprecht 2010: 257).[24]

Agamben's theory then suggests that cinema is the recovery of the loss of the bourgeois gestures and that, as such, cinema is neither a neutral nor a stable technology, nor a technology only (whereby it remains unclear throughout this exploration how narrow Agamben defines *cinema*). Hence, the crisis of representation in the early 20th century has led the Western world into a crisis of human gesture, one that is restored through the gesturality of cinema that exhibits the conditions of cinematic montage as 'a sphere of pure means, that is, of the absolute and complete gesturality of human beings' (59). I will show in Chapter Six that tracking bears similarities to the montage process in cinema.

What emerges for Agamben is the political and ethical dimension of cinema as a technology that is pure *mediality* and *gesture*.[25] For him, 'the element of cinema is gesture and not image (…) because cinema has its centre in the gesture and not in the image, it belongs essentially to the realm of ethics and politics (and not simply to that of aesthetics)' (2000: 49; 54). Cinema is not about images, but about bringing what is made into static images back to life. This aliveness is the *mediality* that cinema montage embodies. According to Gronstad & Gustafsson (2014):

---

[24] According to this thesis, the development is noticeable in the entanglement of hypnosis, hysteria, and photography in the 19th century (Didi-Huberman, 2004).

[25] 'Nothing is more misleading for an understanding of gesture, therefore, than representing, on the one hand, a sphere of means as addressing a goal (for example, marching seen as a means of moving the body from point A to point B) and, on the other hand, a separate and superior sphere of gesture as a movement that has its end in itself (for example, dance seen as an aesthetic dimension). (…) *The gesture is the exhibition of a mediality: it is the process of making a means visible as such'* (Agamben, 2000: 57).

> 'Agamben's main concern, however, is montage and what he calls its "transcendental conditions," which are *repetition* and *stoppage*. Drawing on philosophers such as Kierkegaard, Nietzsche, Heidegger and Deleuze, he points out that repetition is not about the return of the same but rather the return of "the possibility of what was"' (3).

For Agamben, the capacities of cinema technology of repetition and montage are not instrumental or neutral (nor unethical); they also blur the individual components in their making (Agamben does not refer to human agenthood as stable or distinct). For him, the unfolding (what Barad might call the *performative* aspect) is what makes cinema *ethical,* because of its gestural relationality that 'allows the emergence of the *be-ing-in-a-medium* of human beings and thus it opens the ethical dimension for them [corporal movements]' (57). This comes close to Nancy's work on the cinema of evidence, according to Gronstad & Gustafsson (2014), who write that both theorists' work does not orientate itself around appearance and aesthetics, but toward its *unfolding*. For both, this coming into presence of the world, its continuous disappearing and reappearing, holds a profound political importance. In Agamben's words: '(…) the task of politics is to return appearance itself to appearance, to cause appearance itself to appear' (5).

Consequently, early cinema forms a *dispositif*[26] in the transition of biopolitical relations, in which human communicability (as an openness to communicating with others) in the form of gestures is caught in the act of its own disappearance. Yet, if gesture is the site of a potential within cinema to operate historically, it is also the locus of a biopolitical process that manifested in the human body towards the end of the 19th century. Levitt (2011) explains the ethical and biopolitical as an

---

[26] In *What is an Apparatus?* (2009), Agamben reflects on Foucault's term *dispositif* by writing that: 'The term "apparatus" designates that in which and through which one realizes a pure activity of governance devoid of any foundation in being. This is the reason why apparatuses must always imply a process of subjectification. That is to say, they must produce their subject' (11). (*Apparatus* is used as the English translation for his original term *dispositivo* from the original version in Italian)

> '[A]ppropriation of gestures as images, as forms of knowledge deployed in the discipline of bodies, is centrally implicated in the emergence, by the early twentieth century, of a distinctly modern variant of biopolitics – a situation that would reach its horrific apotheosis only a few years later' (2011: 199).

Agamben's abstract and historically interwoven argumentation should not be misunderstood as a literal move beyond human gestures. He is interested in another form of gesture or *gesturality* as a process. Despite humans being still able to communicate in coded gestures, what Agamben wants to address is that the very process of gesture-making as a 'the process of making a means visible as such' (2000: 57) is transformed through cinema. He draws our attention to the initial reading and expressing of gestures being a private matter of the bourgeois individual, which has moved into a public domain as being recovered *within* the new technology of cinema. This step, for him, has generated a reciprocity in which the public domain then penetrates and operates *within* the private body, by mediating its privateness back – but already transformed – into a public realm. Hence, human and technology, private and public, both are, in this case, intertwined agencies, but never separate entities within a reciprocal structure, without an end and/or a beginning and beyond fixed borders or bodies.[27]

What I conclude from both ethical perspectives, the Posthumanist's and Agamben's, is that moral and ethical views on technology can differ substantially and produce, co-shape, and influence different agency models, which then implicate their wider ethics. Furthermore, expressivity of human agency or its meaning is not a one-way street; it is formed by technology, and robots do challenge this formation in particular, by hinting

[27] Agamben's work on ethics and early cinema is widely discussed by Gronstad & Gustafsson (2014), Väliaho (2010), Harbord (2016) and Ruprecht (2017).

to a literal and humanoid simulation of human agency, while also embodying another level of invisible and machinic agencies.

It is important to keep in mind that morality is mostly used as an evaluative concept *of* technology or human agency, while ethics can be the structure of morality evaluations or a completely amoral concept to understand agency entanglement *through* technology. Discussions on morality and agenthood are continued and contextualised in more detail in Chapters Four and Seven. Especially, in Chapter Seven, new concerns and limitations on this term appear as I reflect on why accountability is removed from the agency of technology. Chapter Seven debates this deficit and highlights why the contemporary discussion on morality and accountable agents is drifting into a theoretical (and interesting) but increasingly impractical view on technology.

2. Social Robots Between Humanoid Companions and Monitoring Tools

Chapter summary: Next, I enter the (i) Robot Ethics framework to discuss the ethical debates on the use of social robots in elderly care. I focus on the implications when looking at social robots as companions and how this creates a problematic and anthropomorphic view on their agency. Such anthropomorphic status is given to social robots on the grounds of their humanoid bodies suggesting agenthood due to resemblance, which leads to the anthropomorphic fallacy of expecting human-like behaviour and intentions from them. I specifically survey literature in Robot Ethics and point to the problematic position of the social robot as a companion (as pseudo-agent) or as a monitoring device (as a misused tool), while both emerge from a moralisation of the social robot as an agent-like device. This chapter looks at the work of Lin et al. (2012), Royakkers & van Est (2016), Turkle (2005, 2011), and Sharkey & Sharkey (2010).

The importance of this investigation grounds on the increasing use of social robots[28] in elderly care (Royakkers & van Est, 2016; Lin et al., 2012; Sharkey & Sharkey, 2010; Wu et al., 2010).[29] One concern I have is around the position with which social robots are mainly associated: as a *companion,* human-like (pseudo) agent, or *caring* partner, while their capacities as computational machines is often overlooked or neutralised and not fully examined ethically.

This companionship position is accompanied by various fears and expectations, as Lin et al. (2011, 2012) point out, which means it is not endorsed necessarily. Especially, as some ethicists worry that the social robot is positioned increasingly as a *partner-like* agent, it hereby enters a new discussion on how much moral accountability and responsibility it should be expected to have. Yet, robots still lack any abilities required to be worthy of an agenthood position in the first place, so how can these attributes be thought together? Lin et al. (2011) ask, therefore:

> 'Is it ethically permissible to abrogate responsibility for our elderly and children to machines that seem to be a poor substitute for human companionship (but perhaps better than no—or abusive—companionship)? Will robotic companionship (that could replace human or animal companionship) for other purposes, such as drinking buddies, pets, other forms of entertainment, or sex, be morally problematic?' (945).

---

[28] '*Personal care and companions:* Robots are increasingly used to care for the elderly and children, such as RI-MAN, PaPeRo, and CareBot. PALRO, QRIO, and other edutainment robots mentioned above can also provide companionship' (Lin et al., 2011: 944). I am concerned in this thesis with the wider concept of social robots and less with one specific model.

[29] 'According to the European Commission (2012), the proportion of those aged 65 years and over is projected to rise from 17% in 2010 to 30% in 2060, with the peak occurring around 2040. (…) In Japan, the country with the highest proportion of elderly citizens, the population is also rapidly aging; 23% of the population was already older than 65 years in 2010, predicted to rise to 31% by 2030, and in the United States, 13% were over the age of 65 in 2009, expected to rise to around 19% by 2030' (Royakkers & van Est, 2016: 62).

These questions are critical; however, they are also anthropomorphic, anthropocentric, and moralising, since Lin et al. (2011) expect that robots should judge abusiveness or friendship like human agents supposedly can. They imply that the social robot could have a cognitive ability to be either a *friend* but, in a negative way, to also become a harming *abuser*. Still, these fears are projections of qualities that a robot does not yet own, and, thus, the status of robots as morally relevant agents is projective or even misunderstood. Further, only agents could take a position as a friend or as an 'abuser' and be morally accountable for their actions.

Humanoid Robotics (HR) advocates for the use of social robots being an improvement on the life of the elderly, but Sparrow & Sparrow (2006) state that 'the use of care robots is unethical' (193) in principle. They critique the use of robots by saying:

> 'We see the idea that we can solve the 'problem' of caring for an ageing population, by employing robots to do it, as essentially continuous with a number of other attitudes and social practices which evidence a profound disrespect for older persons' (143).

What 'disrespect' means in this instance is not clear, but it could reflect on the various issues that I raise next. Compared to robot ethicists, roboticists developing and designing social robots are more optimistic about the use of social (or care) companions and are less concerned about ethical problems (as non-ethically motivated research around elderly care shows).

For some, robotic technology will clearly improve the lives and care of elderly patients (Draper & Sorell, 2017; Roger et al., 2012; Chu et al., 2017). In the context of assistance and independence, Chu et al. (2017) believe that a robot can guarantee the medication scheduling of an elderly patient; support the memorising of tasks or even eating schedules of a demented patient, or client; and help with physically demanding

tasks, such as lifting of an elderly person (Draper & Sorell, 2017). In particular, the value of social robots for dementia patients is also positively discussed (Roger et al., 2012; Chu et al., 2017). I do not argue against these positive aspects in this thesis, for it has a different focus of exploration.

Even if several robot ethicists observe social robots critically within elderly care, some are specifically critical of the robot's position as a perceived agent and companion, but not in the way I am. Turkle (2005, 2011), for instance, focusses on the implications of the potential for deception of the elderly. Turkle points to the danger of anthropomorphic projection, which I critique as well.[30] However, Turkle's (2011) caution around the use of social robots is on them as deceptive devices, a view she developed out of her work as an anthropologist at the Massachusetts Institute of Technology (MIT) (ix). She objects to the use of robots for the elderly and for children by saying: 'Roboticists make the case that the elderly need a companion robot because of a lack of human resources. Almost by definition, they say, robots will make things better' (24). For her, the robotics industry is the major ethical problem, since it is the main driver behind promoting these devices.[31]

Turkle (2005) grounds her criticism upon two important facets of social robots. The first is the extent to which they are 'relational artefacts', through which the anthropomorphising design of social robots encourages them to be seen as 'artefacts

---

[30] 'Anthropomorphic projections do not require, nor necessarily imply, the belief that a non-human animal or object has mental states similar to ours. Nonetheless, in many cases, they will lead to the formation of such beliefs, which may or may not be true. Historically, the term 'anthropomorphism' has been reserved to refer to when the attribution fails, and the belief is false' (Damiano & Dumouchel, 2018: 6).

[31] 'According to a survey from World Robotics (2012), sales of service robots for personal and domestic use increased by 19 % in 2011 to 2.5 million units, with projections for 2012-2015 showing increases in domestic, entertainment, leisure and handicap assistance robotics to 15.6 million units overall' (Ford, 2014: 28).

that have inner states of mind' and to interact with them involves an 'understanding these states of mind' (Turkle, 2005: 62). Second, she refers to ethnographic studies in which children and the elderly bond with and care about these 'evocative artefacts' easily and, hence, enter deceptive relationships. The 'illusion of relationship' (Turkle, 2011: 514), is the most damning part for her, since it exploits our 'Darwinian buttons', making us anthropomorphise objects by projecting human-like qualities onto them.

Hence, for Turkle, the main ethical issues stem from social robots being a deceptive technology, one that is pretending to be what it is not; namely, a companion and agent. Even if I am careful in arguing that the bonding process with technologies or devices must be *less* real or authentic than relationships with humans and animals, the critique of a simulated companionship seems valid.

> 'Thus, Turkle sees in social robots a further step in the development of our "culture of simulation," which threatens to turn people away from "real" social relationships – that is, from relationships with other humans – and reduce their social life to an illusion – to the feeling of being together with someone, when in fact one is alone' (Damiano & Dumouchel, 2018: 4).

Deceptive and masquerading companionship would be the opposite of what Sparrow (2015) claims to be the purpose of social robots in elderly care. For him, robotic design for elderly users should be geared towards promoting happiness rather than to achieve seemingly objective measures of welfare (Sparrow 2015). For Sparrow, this means finding a balance between autonomy, privacy, and independence that can guarantee welfare, even if it remains unclear how and who is negotiating these terms.

Sharkey & Sharkey (2010) raise concerns on social robots in elderly care as critically as Turkle does, but differently. They define three dimensions that, although problematic, are worth discussing: (1) Assistance, (2) Monitoring, and (3) Companionship. Whereas

I am more concerned with humanoid social robots, they also focus upon animal-like companionship robots, such as robot seals (Paro) and robotic cats, which are anthropomorphised as well. Sharkey & Sharkey (2010) agree only partially with Turkle's work and how the elderly might be potentially deceived into believing that social robots are companions. 'Essentially, our suggestion here is that to claim that robot companions are unethical because their effectiveness depends on deception (…) [and] oversimplifies the issue' (36). Instead, Sharkey & Sharkey (2010) speak of…

> '[…]two different bases for the associated ethical concerns: human rights, and shared human values. We shall outline these in turn. An emphasis on human rights provides support for the assumption that the physical and the psychological welfare of the elderly is as important as the welfare of others' (2010: 27).

They summarise further potential concerns around social robots as:

> '(1) the potential reduction in the amount of human contact; (2) an increase in the feelings of objectification and loss of control; (3) a loss of privacy; (4) a loss of personal liberty; (5) deception and infantilisation; (6) the circumstances in which elderly people should be allowed to control robots' (27).

Sharkey & Sharkey's biggest concern is, in my view, that the use of robots for care could violate the basic human rights of the person being cared for.[32] Even if their reflection considers the wider consequences of using social robots, they do not assign any agent or moral agency to them. Sharkey & Sharkey position their critique on the opposite spectrum of Turkle's, by viewing social robots rather as *misused* instruments and isolating machines, not as deceptive pseudo-agents. I only agree partially with their

---

[32] 'Depriving senior citizens of social interaction with their fellow humans is an ethical issue that is not explicitly addressed by human rights legislation. Such a right is perhaps implied such as in Article 5 of the Universal Declaration of Human Rights, ''No one shall be subjected to torture or to cruel, inhuman or degrading treatment or punishment.'', or Article 9, ''No one shall be subjected to arbitrary arrest, detention or exile''' (Sharkey & Sharkey, 2010: 29).

critique, since I disagree with them not recognising that social robots have a complex agency nonetheless, despite not being human. What concerns Sharkey & Sharkey is the deprivation of human contact as the elderly become increasingly isolated, objectified, or lose control over their lives and environment. (29) 'The worry is that the use of robots in elderly care for tasks such as lifting, carrying, or even cleaning, might result in a reduction in the amount of human social contact that an elderly person experiences' (Sharkey & Sharkey, 2010: 29).

Their critique denies the agenthood of robots in the first place, since the humanoid design is, for Sharkey & Sharkey, only a superficial and deceptive projection and cannot be taken seriously. For them, the elderly are exposed to a deceptive machine that cannot be considered a substitute for human company at all, which might be true, but it does not make it less influential or more neutral, as I argue. Even if I share their critical point on this exposure being potential isolating, I do not agree that a social robot's major flaw is that it is not an actual human agent and, therefore, must be isolating. Nonetheless, I see a problem in the elderly mistaking the robot's actions as being motivated by a moral goodwill or bad intention and their inability to distinguish between what the robot accounts for or is capable of.

Amanda Sharkey's more recent work (2014) emphasises another problem by expanding the ethical issues around agenthood in social robots. Her research addresses questions on dignity in elderly care. Sharkey (2014) refers to the concerns raised at the National Pensioners Convention in 2012. Alarmed by a report surfacing on how elderly patients were neglected by human caretakers (64), the convention members were urged for a new Dignity Code to protect the care conditions of the elderly patients. The report found out that professional human care staff struggle to keep up with the workload and that

this leads to an increasing neglect of elderly patients. To tackle this concern, the members decided to introduce 'dignity ambassadors' (64) in professional care, which shows a certain awareness around further challenges in the care profession.

Preserving the dignity of the elderly is a goal that must be addressed holistically and beyond the question of dignity, but also as aligned to the exposure to robots as companions or isolating machines (remaining an important concern). Expecting that a robot would understand the concept of dignity is questionable, just as much as the question on how a 'dignity ambassador' could be able to intervene and identify a (human?) caretaker that treats the elderly without dignity.

However, for Sharkey, there are two sides to this scenario. It might be that the introduction of robots could improve this worrying situation by providing support for tired and overworked care-givers. She argues, first, that it could. A robot might be (perceived as) 'kinder' (65) by a patient and, therefore, more positive than an inattentive or unkind human carer. A robot could be perceived as less judgmental when helping the elderly with their personal hygiene, for instance, and, as robots gain in verbal and interactional sophistication (one day), they might in the end become *better* companions than human caretakers.

However, Sharkey (2014) clarifies that to counter such suggestions…

> '[…]it should be pointed out that contemporary robots are poor substitutes for human company. Robots may not exhibit the worst sides of human behaviour, but neither are they capable of real compassion and empathy or understanding' (64).

For her, in the context of the current development of robotics, a preponderance of social robots taking care of older people would deprive these people of human companionship,

and this is what would make many peoples' 'lives (...) unacceptably impoverished' (65). Sharkey (2014) explains that the social isolation following from this preponderance would also have health implications.

> 'For example, being single and living alone has been shown to be a risk factor for dementia (and a) decreased social engagement from midlife to late life was associated with an increased risk of dementia' (65).

What emerges from Sharkey's research is that ethical consequences arise independently of the intention of the industry, or of the developer, be it potential isolation or even physical harm of the elderly due to the misuse or malfunction of the robot.

There is an additional and important ethical facet to social robots leaving the companion perspective. Such refers to the social robot's capacity to monitor the elderly with its embedded cameras or interactive sensory channels (Chapter Six will expand on the computational interaction further). As I am leaving the discussion on the companionship and a pseudo-agency position behind, the intrusion of people's privacy emerges as a new concern in Robot Ethics.[33] Such concern is underestimated in my view and is understood insufficiently, compared to the discussion I will offer in Chapter Three led by (ii) Media and Surveillance Studies.

I want to remain with Robot Ethics for now to outline that this framework is aware of the concerning monitoring capacities of robots, but robots are rather viewed as *spying* tools used by immoral human agents. I argue that this is a projective concern, since the

---

[33] The REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THECOUNCIL of 27 April 2016 regulates: '(1) The protection of natural persons in relation to the processing of personal data is a fundamental right.' And further points to '(3) Directive 95/46/EC of the European Parliament and of the Council (4) seeks to harmonise the protection of fundamental rights and freedoms of natural persons in respect of processing activities (…).' Available at https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN (Accessed 15.05.2018). How the right on personal data will be *harmonised with the processing activities* is legally challenging in my view.

intention to survey – to be immoral – is lacking in the robot; hence, the ground for a moral agency is lacking too. However, I highlight next that the very ability to interact and monitor is still an ethical process in which various agencies align to become potentially problematic if they are not fully understood. The problem I see within Robot Ethics discussing monitoring as an issue is that it remains unclear if the monitoring and privacy intrusion is meant to be an *immoral* intention drawn from the robot's (perceived) immoral agenthood, or from the intentional misuse of the robot by an immoral human agent that spies on the elderly.

Monitoring leads to the concern of privacy intrusion, which is mentioned (even if rarely) by Royakkers & van Est (2016). They identify that the ability to monitor and survey the environment is an ethical concern when using robots. They state:

> 'Robotics can be applied in all sorts of ways to monitor certain situations, such as the patient's state of health, a car driver's focus of attention, and the safety situation in the street or on the battlefield. (…) The utilization of robotics calls for an explicit and careful consideration of various potential conflicting interests in various fields, for example, between health and privacy, (…). It should also be taken into account that the utilization of this kind of information technology may go hand in hand with more intensively keeping on the actions of caregivers, (professional) drivers, police officers, and/ or soldiers' (310, 311).

They dedicate a substantial part of their philosophical work to the use of social robots, and other robots such as 'domotics'[34], which are also used in elderly care. However, their critique on privacy intrusion is brief when it comes to elderly care and monitoring, since, for them, privacy intrusion through monitoring, is limited to long distance care services.

[34] Domotics are, for instance, supervised 'devices and infrastructures in and around homes that provide electronic information for measuring, programming, and controlling functions for the benefit of residents and the providers of services' (Royakkers & van Est, 2016: 95).

Although Royakkers & Van Est's confrontation with these ethical concerns is, in my view, too brief and superficial, their research at least addresses privacy issues in the context of social robots as monitoring devices. As I just argued, to position the problem of monitoring as one of immoral intention, this requires for the robot to have, first, intention and, second, immorality; both attributes, which are anthropomorphised and projected into robots at this point.

The second aspect is that the concern on privacy intrusion is often disconnected from social robots - it is mainly brought up as attached to the ethical discussions around drones (2016: 160-166). This suggests to me that Robot Ethics is not only limiting its ethical discussion through the focus on the humanoid shape of the robot, but also by the context of its use. I point to this deficit in Chapter Seven when arguing that social robots must be treated always as digital technologies with embedded capacities to track or collect data, despite their shape and use.

Sharkey & Sharkey (2010) also raise a critical awareness on monitoring and privacy, even if not a sufficient one in my view. For them, the monitoring ability of the robot has positive and negative elements.

> 'Robotic surveillance devices have already been developed for warfare, for policing, and for home security (Sharkey 2009) and these could easily be adapted for monitoring the elderly. A robot that traverses the house, and relays information picked up by its sensors, is something that is well within the current technological limits' (2010: 32).

One advantage of monitoring systems is, for them, that it enables virtual doctor visits in cases in which the patients live far away from their medical support. 'A monitoring robot could increase the safety of an elderly person in their own home, and make it possible for medical staff to virtually visit the elderly person and provide health checks'

(2010: 32). However, this ability to monitor and do virtual examinations links to their initial fear of the elderly being isolated and having even less human contact (32).

Two issues remain unresolved for Sharkey & Sharkey in the debate about monitoring: the elderly patients giving consent to be monitored, and the access to this information. They argue that this is specifically problematic if a patient with Alzheimer's, for example, has been deemed as lacking capacity to consent, in which case, he or she would probably

> 'forget that the robot was monitoring them, and could perform acts or say things thinking that they are in the privacy of their own home. Moreover, who should have access to the information, and how long it should be kept for? With the massive memory hard drives available today, it would be possible to record the entire remainder of an elderly person's life, but this is not something that they would necessarily consent to if they were able to' (2010: 32).

They point to another advantage of monitoring: allowing reach to demented people through the development of smart homes for the care of dementia sufferers, which then become useful technologies able to determine if an elderly person has fallen over or needs support (271). The concerns they have around the lack of consent relates to the critique I raise as well, which is on the lack of media literacy provided to the people affected by these new devices.

Draper & Sorrell (2017) have studied how much autonomy or independence elderly patients accept or encourage in a robot. Therefore, in my view, achieving a holistic media literacy of those exposed to robots must be an essential driver for future research, especially as the increased 'mediatisation' of social structures (Hepp & Krotz 2014; Krotz, 2012; Lundby 2014) influences not only elderly care, but affects various

environments.[35] Obviously, if the ethical discussion in Robot Ethics is limited or incomplete, the literacy on the technology in place will lack important aspects, as I am increasingly highlighting. Elderly care and Robotics are equally affected by the convergence of Information and Communications Technology (ICT) structures and the implementation of new digital technology.[36]

Hence, the study by Wu et al. (2011) seems worth discussing, in which they surveyed the opinions of a group of elderly care clients on humanoid and social robots.[37] What the researchers did was to engage the elderly with humanoid social robots, all of which exhibited a variety of human-like features, shapes, and likenesses, from very human-like to less human-like. Most responses were not in favour of the robot exhibiting a perceived human visual alikeness. Wu et al. (2011) explain:

> 'Most of the participants expressed a strong reluctance (they often use the term 'fear') toward a robot conceived as a substitute of a human presence. They often evoked some social and political issues, such as dehumanisation of our society. ''I cannot imagine a world ... with no contact, no one to speak to you'', said a participant' (124).

It was not just fear (or a sense of uncanniness) towards a robot as a pseudo-human device that was brought up, but also, as Sharkey & Sharkey and Turkle said, the fear of isolation. Furthermore, what their findings indicate is that, although participants were

---

[35] 'In a 'media age', mediatisation is the concept that would 'acknowledge media as an irreducible dimension of all social processes. (…) This, further, refers to communication as part of all social processes, to how media shape communication processes' (Couldry, 2012: 136, 137).

[36] Royakkers & van Est (2016) claim that 'the modern robot is not usually a self-sufficient system. In order to understand the possibilities and impossibilities of the new robotics, it is important to realize that the service robot is usually supported by a network of information technologies, as is, for example, the Internet. Thus, this implies, in particular, networked robots' (11).

[37] 'A total of 15 older adults over the age of 65 (range from 66 to 89 years old) participated in three focus group sessions (4 in the first, 5 in the second and 6 in the third). Three participants were male and 12 were female. Thirteen of them were recruited from the Memory Clinic of the Broca Hospital and two were recruited from an association for the elderly' (Wu et al., 2011: 122).

hesitant toward some humanoid robots, they did show very positive attitudes toward smaller robots with human traits.

> 'Whether a robot has a high degree of human likeness did not matter for them, instead, they seemed to be attracted by the kind of humanoid robots which somehow look different from human-beings and which are creative' (124).

Interestingly, Wu et al. assess their own findings about humanoid appearances to be consistent with other studies, such as the one undertaken by Dario et al. (1999) who had shown previously that anthropomorphic robots were less socially acceptable when compared to machine-like robots. The final prototype was 'anthropomorphised', but 'also still looked like a machine' (124).

I suggest, however, that we should view these results with a certain amount of caution. What the elderly exhibited as reluctance could be due to their lack of familiarity with robots or new technology in general. This scepticism could disappear through time and use. Wu et al. (2011) do not address this angle. However, what they highlight further was that the elderly participants had an interest in the robotic capacities beyond their aesthetic shape. This group of elderly people was not naive or ignorant in asking about the robots' qualities as machines.

> 'Beyond aesthetics, participants questioned in fact the values underlying the design of each type of robot, or in other terms: "what do roboticists have in mind when designing this type of robot?'' Some anthropomorphic robots were challenged with this question: ''is it ok to copy human beings?'' For those who do not like the idea, some robots are appreciated simply because they do not pretend to look like human-beings' (125).

The fact that social robots are tracking devices does not only raise a naïve interest, but it does complicate the discussion on their agency enormously, as I have hinted to various

times. It will prove that the association of monitoring to tracking here, and to dataveillance in the next chapter, goes beyond what Robot Ethics considers within their epistemological explorations of potential ethical issues. The problems on dataveillance escape the morality and agenthood framework, since these require a different angle on technology and a broader consideration of ethical structures. Bringing up monitoring superficially, as happens in Robot Ethics, stems from not fully understanding the dimensions of data concerns and tracking. Elderly people being denied more literacy on data and computational interaction of robots is, therefore, only a consequence of the initial misjudging by Robot Ethics.

To address these deficits, I move to exploring the inherently ethical dimension in social robots by focussing on their ability to collect and process data, which I see as preceding monitoring discussions. The reason for doing this lies in the necessity to understand the increasing sophistication of tracking systems as 'advancing biometrics capabilities and sensors, and database integrations' (Lin et al., 2011: 946), which enable better monitoring, leading to an intrusion of the private lives of elderly clients or patients.

To understand monitoring, tracking must be explained as a fundamental characteristic of a social robot's interactivity. I refer to tracking as a multi-layered process that is embedded in the robot through algorithmic modules and software, and is part of what Read (2014) calls the 'inner shell'. Most social robots (and other responsive robots) are fitted with tracking modules or systems and embedded camera systems that enable them to detect, as part of the Human-Robot-Interaction (HRI). The social robot must be understood beyond its humanoid shape as a complex of multiple networks and

technologies (Royakkers & van Est, 2016: 97-99)[38], in which digital tracking modules are embedded as algorithmic patterns to process data. In fact, tracking is already data and data-processing. Data-processing and data-gathering create the ethical ground for this discussion, since, on the one hand, the robot cannot respond without data input and, on the other, what happens to this data is not made transparent or controllable.

Tracking is extremely difficult to perform as a process; what is even harder is its synchronisation to locomotion and movement (Brèthes et al., 2004; Rossini, 2012). For all of these elements to come together, two things are needed. First, the robot's tracking module requires clear instructions and accurate concepts of what to detect (human, gesture, emotion). These are programmed by a human developer into the tracking module. Secondly, it relies on the computational capacities of the robot to be highly autonomous and to process information in real-time; otherwise, no response or interaction with the human subject – no Human-Robot-Interaction (HRI) – is possible. The most common input channel is a visual camera system.

As early as 2002, the research on social robots by Fong et al. (2002) pointed to tracking as an important social capacity of social robots. Interestingly, one of the first abilities embedded into robots, while still being used in factories, was vision. 'Robot technology progressively masters more and more complex operations. This is made possible by improved visibility (via 3D vision systems), better navigation and mobility, better voice recognition, and smarter interaction with people' (Royakkers & van Est, 2016: 3).

[38] I avoid limiting the exploration to one specific robot, since I offer a wider philosophical reflection on social robots as companions and why such view collides with the position of social robots as tracking devices. I consider robots like *Pepper* or *NAO* to be illustrative for a social robot.

What I emphasise is that, by introducing tracking, the question on data collection becomes ethically charged and structurally inherent to designing and using robots. This must therefore be raised much more vehemently by Robot Ethics than has been done so far. As such, it is not sufficient to address privacy intrusion as a potential ethical issue only in the case of data infringement or immoral monitoring being proven, since the latter is unlikely, considering no roboticist would admit doing this.

To conclude, positioning the social robot as a companion creates projective and limited views on its agency by moralising its appearance (the anthropomorphic concern), but viewing it as a monitoring machine is moralising the tracking quality as misused function (the anthropocentric concern). Both concerns are ontologically flawed, since they require social robots to be as conscious as humans, or to understand morally *bad* decisions so they act as moral agents (in terms of them intentionally harming people) (Dodig Crnkovic & Cürüklü, 2012; Coeckelbergh, 2010). On the other hand, the robot might not be considered the spying agent itself (again, depending on how much agency is projected into it and by whom), but as being a misplaced and misused spying tool only by the human agent taking the immoral position instead. This presents the social robot as an immorally applied instrument without much agency either.

Reviewing Robot Ethics literature showed that is it unlikely to identify obvious immoral or unethical behaviour in the robot or the developer. But nevertheless, it is concerning to place social robots with one use in mind (companionship) that might have several and underestimated side effects (data gathering), especially if the effects become a new goal. Since, there is no guarantee or clarity on what the data gathered by social robots is used for and who decided on this, social robots are affecting people's privacy and, indeed, compromise their privacy.

The previous exploration strengthens my initial hypothesis: that Robot Ethics is morality-led in its understanding of robotic agency and that this view not only anthropomorphises robots, but does not look deep enough into their wider influence or consequential use, ethically. As I reflected on the companion (robot as friend and pseudo-agent) and the monitoring (robot as monitoring device intruding on privacy but not being an agent), their hybrid agency models are not clearly understood in this framework. Also, the ethical concerns commence too late as a discussion around misuse and harm, instead of supporting a discussion inherently built on the social robot's capacities to interact. Consequently, the ethical dimension of data is worth discussing in-depth next.

3. On Ethical Issues Around Dataveillance and Social Robots

Chapter summary: This chapter builds the antithesis to the previous concerns on social robots as companions or monitoring devices, and zooms into the robot's ability to track and the wider ethical implication of this process. I avoid evaluating the human intention to misuse technology or to intrude upon someone's privacy. Instead, I focus on an early intention; on the very gathering of data in sensitive contexts as an ethical concern. I make use of Media and Surveillance Studies (MSS) canons in order to illustrate that the ethical discussion in Chapter Two has given us a limited understanding on why monitoring is ethically problematic. Instead, I highlight how the ethical problems of social robots begin with their ability to gather data and their subsequent placement in elderly care, which is not a neutral placement of their human creators. I pick up the failure of Robot Ethics to look at the digital tracking ability from a data perspective and debate why and how dataveillance as a structure creates a new form of thinking about surveillance and monitoring. This discussion is driven data-related MSS research from Gitelman (2013), Andrejevic (2012), and Ball et al. (2012).

For Royakkers & van Est (2016), Lin et al. (2011, 2012), and Sharkey & Sharkey (2010), the focus on ethical issues arising from the use of social robots is placed around perceiving the social robot as a companion, or on the misuse of the social robot by a human agent. These views are anthropomorphic and anthropocentric and do not consider the aspects of tracking or data gathering as inherent to the robot's ability to interact. Robot Ethics also does not fully consider that the making and collection of data is an inherently non-neutral process and ethically problematic.[39]

However, Robot Ethics does address tracking modules as an important function for the social robot. It acknowledges that this function is open to misuse because it enables the monitoring of people and this could *potentially* intrude upon their privacy (Royakkers & van Est, 2016). What Robot Ethics and HR do not acknowledge is that the ethical issues around digital tracking technology are intimately linked to social robots. Stahl & Coeckelbergh (2016) draw upon a techno-philosophical angle to identify this as a problematic area. They argue that the ethical issues of social robots must be re-thought when using what they call 'healthcare robots', while also claiming that the 'a priori' agenda in ethics is crucial (154).

According to Stahl & Coeckelbergh's (2016), Robot Ethics is an area that does not reach out enough to other fields or actors, such as to health care providers (154), to expand its ethical views around robots and care. The discussions that emerge in Robot Ethics are therefore, 'located "in the head" of the philosopher-developer' (154), not in the practical field. In my view, they correctly address the lack of inclusion of other

[39] The Federal Trade Commission Brokers Report (2014) states that 'new forms of tracking and increasingly powerful analytic capabilities have emerged, such as mobile tracking and analytic services that enable tracking of users across devices' (5). Such are enabled by social media websites and mobile applications and have 'dramatically increased the availability, variety, and volume of consumer data' (5).

shareholders in this debate, but neither do they include the Data Ethics angle from Floridi & Taddeo (2016).[40] Neither do Stahl & Coeckelbergh address dataveillance as a concern. Further, Robot Ethics is not aware that its epistemological framework of discussing tracking is limited to the Human-Robot-Interaction (HRI) and movement detection, while the HRI framework only provides a *partial* definition of tracking, one that is non-ethical. I consider this a huge deficit and, hence, see my contribution in specifically linking between Robot Ethics, HR, social robots, tracking, and MSS.

As I explore the inherent concerns with data (its gathering and management), I must leave (i) Robot Ethics to enter (ii) Media and Surveillance Studies (MSS) (Schermer, 2007; Kroener & Neyland, 2012: 145), and to examine the non-robotic angles of POT streams so I can find answers on how to understand data collection holistically. Again, tracking in both frameworks is a synchronising process embedded as robots' capacity on making, collecting, and managing data (Ball et al., 2012; Raley, 2013). I introduce the shift from intentional and strategic monitoring[41] to dataveillance by giving an overview on how to understand *data* ethically. I will support a view on data as incorporating *intentions* already, aligned to Gitelman's (2013) work. This does not mean that the social robot cannot still be used as a device spying *purposefully* on the

---

[40] For Floridi & Taddeo (2016), Data Ethics is an important field emerging from Information Ethics that deals with moral and ethical challenges related to the 'extensive use of increasingly more data— often personal, if not sensitive (big data)—and the growing reliance on algorithms to analyse them in order to shape choices and to make decisions (including machine learning, artificial intelligence and robotics), as well as the gradual reduction of human involvement or even oversight over many automatic processes, pose pressing issues of fairness, responsibility and respect of human rights, among others' (2). The association of Robot Ethics to Data Ethics is brought up on a side note and cannot be fully explored in this thesis.

[41] The terms *monitoring* and *surveillance* are used interchangeably in Robot Ethics, whereby, as I mentioned already on monitoring in Chapter One, surveillance as an ethical problem is brought up in the context of drones (Royakkers & van Est, 2016: 131 ff.), rather than around social robots. Sharkey & Sharkey (2010) claim that: 'Robotic surveillance devices have already been developed for warfare, for policing and for home security (Sharkey 2009) and these could easily be adapted for monitoring the elderly' (32). They mainly refer to monitoring being a problem, less than surveillance.

elderly, but it implies that such an intention is not required for the ethical issues to emerge.[42] I will explain that intention and ethical relations surface when the decision on data gathering is made, when modules are programmed, and when the robot is placed into the care environment, which are three neglected moments in Robot Ethics.

According to the media and MSS researcher Galloway, the term data originates from the Latin *data*, which means 'the things having been given' (Galloway, 2012: 82). The Merriam Webster dictionary has a much more pragmatic definition of data as 'factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation'.[43] The interesting aspect of data, which makes it so hard to grasp and imagine, is that it is abstract and produced in a non-image form; even if Galloway (2012) argues that it has a 'phenomenological claim' (82), it is not yet information.

Gitelman (2013), on the other hand, looks at data slightly differently, in pointing to its potential of becoming information. Gitelman's book *"Raw Data" Is an Oxymoron* (2013) exhibits her criticality immediately through its poignant and self-explicatory title appropriating Bowker's statement: '"Raw data" is both an oxymoron and a bad idea' (Gitelman, 2013: 1). As a collection of texts of various authors, the content ranges from the history of data to the idea of 'data friction in the field of astronomy' (8). Gitelman's collection is not historical, but circles around the problematic dimensions of data and

[42] The social robot is considered as a tracking device only in the case that its tracking modules are actively used for the collection of data. This is differentiated from cases in which the social robot is used as a perceived agent, is remote-controlled, and does not gather data, but executes only pre- programmed movements. The reflection of Rossini (2012) on the difference between functional and responsive modules is helpful and is brought in Chapter Eight.

[43] Available at https://www.merriam-webster.com/dictionary/data (Accessed 20.03.2018).

disciplinary boundaries around data science. Its holistic trajectory points to the illusion of data being 'given', 'neutral', or 'objective' (2, 3).

The contributions collected in her book are equally concerned with the 'ethics surrounding the collection and use of today's 'Big Data' as their 'particularly pressing concern" (11). Particularly consequential is her claim that the 'phrase raw data – like jumbo shrimp – has understandable appeal' (3), but is simply incorrect. It makes us believe that 'data are transparent, that information is self-evident, the fundamental stuff of truth itself' (3). She warns that our lack of criticality allows us to ignore the fact that data is always '"collected," "entered," "compiled," "stored," "processed," "mined," and "interpreted"', hence, at 'a certain level the collection and management of data may be said to presuppose interpretation' (3).

Gitelman's biggest issue with scientific and engineering explorations is that they confound simplistic dichotomies like theory/practice and science/society in a rich, diverse body of work that, among other things, has explored the situated, material conditions of knowledge production. Looking at the ways scientific knowledge is 'produced – rather than innocently "discovered," for instance – resembles our project of looking into data or, better, looking *under* data to consider their root assumptions' (4). Her critique is not that the search for objectivity is a 'bad' (4) thing, although that is not possible, considering the epistemological frameworks through which data is managed. Therefore, the obsession for objectivity is an

> 'abnegation, neutrality, or irrelevance of the observing self, [and it] turns out to be of relatively recent vintage. Joanna Picciotto has recently suggested that "the question raised by objectivity is how innocence, traditionally understood to be a state of ignorance, ever came to be associated with epistemological privilege"' (4).

However, the idea of the 'innocent observer' (4) that endorses the assumption similar to the existence of mechanical objectivity, led to objectivity emerging 'as a dominant ideal in the sciences only in the middle of the nineteenth century' (5), and was closely tied to the development of photography during those same years. According to Gitelman, the problem with the assumption that photography was essentially objective is that:

> '[T]he presumptive objectivity of the photographic image, like the presumptive rawness of data, seems necessary somehow— resilient in common parlance, utile in common sense—but it is not sufficient to the epistemic conditions that attend the uses and potential uses of photography' (5).

The history of objectivity is, for her, not fully understood without the history of subjectivity and the creation 'of the self', which is often neglected as a central driver in thinking data. 'Data require our participation. Data need us. Yet for the suggestive parallels, the history of objectivity is not the history of data' (5). While this way of thinking is common for researchers like Gitelman and other authors in MSS, this aspect also has implications for research on machine morality and POT research. The idea of the engineer influencing and shaping the computer code with his or her values will be discussed through Wallach & Allen's work (2009) in Chapter Seven, even if their focus will moralise technology problematically.

Following Gitelman's arguments further, I consider another important aspect worth highlighting, which is that '*data are aggregative'* (5) and 'plural' (8). These aspects are interesting, but most importantly, they allow me to question the search for a single moral agenthood in social robots, which occupied Robot Ethics and the companion position. The shift from a single agenthood to plural agencies, as Floridi and Brey suggest in Chapter Seven, might not be directly linked to Gitelman or to MSS, but it undergoes a similar thought process on the plurality and complexity of data and

algorithms. Therefore, this early realisation on data as a plural concept is helpful in order to build further concepts around data as distributed agencies. The plurality of data is argued by Gitelman as an interesting rhetorical conflation of data between being a singular phenomenon and plural phenomena. According to Gitelman, data piles up and is accumulated into data sets; this aggregation, leading, in its extreme cases, to dataveillance, follows the principle: 'more is better, isn't it?' (8). But what this suggests is in fact a structural shift, not just a growth of the quantity of data. Therefore, the blurriness of data as singular or as plural must be clarified, since the plural is not simply the *addition* of singular data. She declares that:

> '[S]entences that include the phrase "data is . . ." are now roughly four times as common (on the web, at least, and according to Google) as those including "data are . . ." despite countless grammarians out there who will insist that *data* is a plural. [...] Data's odd suspension between the singular and the plural reminds us of what aggregation means' (8).

Gitelman then also amplifies:

> 'The singular *datum* is not the particular in relation to any universal (the elected individual in representative democracy, for example) and the plural *data* is not universal, not generalizable from the singular; it is an aggregation. The power within aggregation is relational, based on potential connections: network, not hierarchy' (8).

This realisation that data are *intentional* and *plural* suggests that any technology operating with digital data does not allow to be understood as a single or stable agent, but should be seen beyond any singular agenthood; even if the humanoid body of the social robot encourages this anthropomorphic fallacy, it would be an incorrect analogy. Reflecting on data as I suggested here transforms the social robot into a relational network of data input and output, and this complicates the discussions in Chapter Two immensely. I do not only offer a different perspective through the data angle, but a

holistic one, while not suggesting to underestimate the perception of robots. I rather urge not to build their ethical agency upon their appearance exclusively.

I argue that the aggregation process of data adds a new ethical dimension to the existing discussions on monitoring in Chapter Two.[44] If ethical structures are understood as relational and reflective engagement and dynamic, as Ward (2015) stated, then the first ethical links between the social robot and the human subject already emerge prior to the application of tracking, at the stage when robots are designed with embedded human concepts and values making use of the algorithmic autonomy.

Next, I would like to continue with the shift from a strategic misuse of technology to the redefinition of surveillance to dataveillance. This is an important step in understanding the ethical issues that digital technologies have introduced, which are very complex to oversee fully. In the *Routledge Handbook of Surveillance Studies* (2012), Ball et al. offer an interdisciplinary collection of contemporary discussions on how surveillance – as a concept but also structure – moves towards the employment of already existing and omnipresent information structures, which allows the gathering of vast amounts of data that can be used in various ways after being collected.

This shift has hugely transformed social structures, since it replaces the old-fashioned and anthropocentric fear of a 'Big Brother', or an invisible governmental agency as the

[44] The ethics relationship between humans and technology is, in this thesis, viewed as an entanglement of human concepts and technological capacities leading to 'ethical complexities' (Braidotti, 2006: 16). This also aligns with Luhmann's view on ethics as a 'reflective theory of morality' (Luhmann, 1989: 360) that 'does not intend to provide guidelines for a practical morality' (112). From their perspective, the moralisation of technology would be an insufficient angle to understand why data issues are ethically problematic.

spying agent.[45] The *new form* of spying has now become a by-product of simply applying a digital and data gathering technology in new contexts, such as elderly care. Ball et al. (2012) emphasise this shift within surveillance structures as one, from a 'strategic surveillance' to a distributed, non-strategic information structure (Ball et al., 2012: xxv). What their publication stresses as well is that the moral intention in using or misusing technology is embedded in the making of digital technologies at their very inception, and not just in their actual use.

This approach points to another factor worth bringing up: the difference between data and Big Data. Even if data has already proved to be ethically loaded with intentions, *dataveillance* is understood as the application of sophisticated digital techniques and technologies, which are used for the manipulation and processing of huge data aggregators (Püschel, 2014, 3; Andrejevic & Burden, 2014). Clarke (1988), who formed the term dataveillance in 1988, defined it as a 'systematic use of personal data systems in the monitoring or investigation of the actions or communications of one or more persons' (499). The fact that dataveillance derives from structures already in place (simultaneously expanded) transforms dataveillance to being an 'algorithmic surveillance' (Introna & Wood, 2014) because of its structural entanglement with the algorithmic and computational level from which it feeds.

Raley (2013) is particularly interested in the implications of dataveillance that go beyond any intentions to *actively* survey. In *Dataveillance and Countervailance* (2013), she gives an insight into the increasing attention paid to how we, as human users or

[45] I have also considered similar tendencies in texts such as Elmer's *Panopticon—Discipline—Control* (2012), in which he provides a valuable discussion on how to understand the surveying agent position. He does this by reflecting on Foucault's and Bentham's concepts of the 'panopticon' and opposes their work to Deleuze's, which 'has tended to lend more weight to networked and immanent forms of surveillance' (Elmer, 2012: 22).

developers, are not only collecting data, but have become the 'resource for data collection that vampirically feeds off of our identities' (10). This destabilising process not only creates ethical issues, because of data gathering becoming inherently linked to how we establish and communicate our identities, but also questions the position of human agenthood as a stable one. This process is captured by Ravetto-Biagioli as the 'digital uncanny' (2016: 3), which, for her, is a moment of uncanniness that humans experience when using (interactive) digital technologies.[46] This process is uncanny because it reinforces an uncertainty and destabilisation of the human agent when encountering the autonomy and agencies of digital technologies (which she discusses through the work of artists such as Lozanno-Hemmer and Viola).

She argues that digital technologies have completely destabilised the human position as an autonomous, independent agent, and that this can be experienced through the interactive processes like visual tracking. For her, these new (digital) forms of 'uncertainties' (2) have shifted debates about where embodiment takes place and have blurred the line between human and technological agency.

Dataveillance is similarly uncanny, since it detaches human agents from being able to control their data ownership or privacy, by transforming agent subjectivity into the new data sets. Compared to how the social robot is viewed in Robot Ethics, dataveillance is not seen as a single or separate technology or entity, but as an abstract, informational and processing structures as Kusnetzky (2010) points out. Instead to dataveillance, he

---

[46] Ravetto-Biagioli's *The digital uncanny and the ghost effects* (2016) was an essential source for my understanding of the computational destabilisation that tracking might lead to. It also introduced a link between digital technology and uncanny humanoid robots, but does not relate these. Ravetto-Biagioli points to the history of uncanny automata herself, but does not examine how digital robots might be digitally uncanny as well. A publication and a seminar presentation emerged from this research (Stamboliev, 2018) in cooperation with Abigail Jackson and can be found in the Appendix.

refers to Big Data as 'a bundle of new methods and technologies for the collection, storage and analysis of vast and randomly expandable amounts of data in volume and of different structures' (Kusnetzky, as quoted in Horvath 2013: 1).[47] Dataveillance or Big Data structures imply further processes as intentional surveillance shifts towards a systematic dataveillance as part of the information structure. Firstly, these indicate the overlap of intention, structure, and contexts in the use of technology. Secondly, they demand an expansion of information structures. And, thirdly, they enable an endless appropriation potential of data after its collection.

The potential of dataveillance will increase, in my view, because of the infiltration of digital technologies into more newly discovered environments. Schermer (2007) states that 'surveillance practice will become more efficient, more user friendly, and more complete through the use of agent technology' (133). He further mentions that this is supported by 'the rapid expansion of surveillance as a result of the use of information and communication technologies' (133).

This expansion is also mentioned by Floridi (2014) and explored by Van den Hoven (2010) as leading to more ethical consequences due to the increasing applications of ICT technology (60-62) in society.[48] What they argue is that the omnipresence paired with the capacities of digital technologies will change the effectiveness, scale, and

---

[47] The original German quote is: 'ein Bündel neu entwickelter Methoden und Technologien, die Erfassung, Speicherung und Analyse eines großen und beliebig erweiterbaren Volumens unterschiedlich strukturierter Daten' (Horvath 2013: 1).

[48] Van den Hoven (2010) says for ICT that these are the 'expression of prior choices, norms, values, and decisions ICT applications are not neutral, but contain the values and norms of those who have designed and engineered them' (6). He further highlights that: 'An abundance of research provides evidence of intentional or inadvertent incorporation of norms in software (…) Finally, ICTs revolve around new entities, such as digital computers, software and information goods, which give rise to new practices and experiences. This makes it sometimes difficult to account for them in terms of traditional moral and legal views' (7).

characteristics of surveillance, which then shifts in both quantitative and qualitative dimensions respectively (Schermer, 2007: 133).

Andrejevic (2012) refers to the dangers of digital technology as grounded on the ubiquity of information technology and locates the ethical problems in the 'ubiquitous' (90) structure of technology, by default becoming an 'ubiquitous surveillance' (90).

> 'Broadly construed, then, the notion of ubiquitous surveillance refers to the prospect of a world in which it becomes increasingly difficult to escape the proliferating technologies for data collection, storage, and sorting—the fact that, as David Lyon puts it, "our whole way of life in the contemporary world is suffused with surveillance"' (Andrejevic, 2012: 90).

This is enabled through the pervasion of communication technologies and the broadening of spatial infiltrations. What this means is that more and more environments are regulated by technologies under the umbrella of 'networked interactivity' (90), which allows these technologies to 'recognize' (90) human movements and actions more broadly, allowing for the technologies to 'recognize us wherever we go, responding to our presence in ways that incorporate information about our histories, desires, needs, and wants' (91). Andrejevic refers to cities and public environments rather than elderly care or social robots, but he illustrates how our daily lives are increasingly infiltrated and mapped through tracking (or sensor) technologies, which might at first create spaces of 'convenience, assistance, and efficiency, but they are also spaces equipped with an unprecedented potential for repression' (91).

There is evidently a danger that elderly care might become another environment into which technology is embedded as an information structure, by expanding the contexts to collect data for commercial or surveillance purposes. Such infiltration of spaces and environments in which technologies permanently gather data as a form of ubiquitous

surveillance also relates to what Gitelman (2013) considers to be embedded in the *plurality* of data. The omnipresence of new technologies as a way to extend and mask surveillance is not a new development; two decades ago, it was Manovich (1996) who compared the operationality of the newly emerging internet with communist surveillance networks.

Laney (2001) also picks up the aspect of ubiquity of digital technologies, in presenting the concept of the 'three V's' of Big Data. He sums up what aspects must come together for dataveillance, or what he calls Big Data, to emerge and to be more than the sum of its parts or the simple accumulation of data. The first V, referring to *Volume,* means that the amounts of data collected are vast due to the variety of sources (all kinds of sensory inputs, including tracking sensory systems) and the variety of technologies. The increased volume of new data being gathered is linked to the omnipresence of technologies enabling it. Andrejevic (2012) refers to this process as *data mining* (as does Schermer, 2007), which emphasises the new possibility to systematically gather vast amounts of data by default. Andrejevic writes that:

> '[I]f the imperative of data mining is to continue to gather more data about everything, its promise is to put this data to work, not necessarily to make sense of it. Indeed, the goal of both data mining and predictive analytics is to generate useful patterns that are far beyond the ability of the human mind to detect or even explain' (74).

The second V refers to *Velocity,* and the possibilities of real-time data streams and the increased processing speed in the gathering and collecting of data in real-time. This also links to the first V – *Volume* – by default, since this aspect enables the gathering of *more* data volume from *more* sources due to *better* processing. This proves specifically important for tracking modules, as I show in Chapter Seven.

The third V refers to *Variability* (and is aligned with Complexity later). Both variability and complexity refer to an additional and crucial difference to former strategic surveillance, since the variety of unstructured data collections is encouraged; the decision of what to make use of can be made afterwards. This aspect relates to the previous two Vs, but highlights the new possibilities in coordinating and structuring data input *posteriori* to its collection. Instead of surveying for a closed or restricted area for information, the useful or valuable information is drawn from a vast amount of data already collected and clustered according to the information needed. The informational level can be decided on the grounds of the patterns that are desired, since the amounts of data gathered (sometimes referred to as a *full take* approach) allow for endless data patterns to be created.

For Kroener & Neyland (2012), this points to the struggle to retain the integrity of the obtained information, or 'footage' (145) (they discuss the context of CCTV cameras), since such footage, once owned, can be utilised and mobilised in multiple ways after its collection:

> 'In the latter case, this raises questions of who or what narrates surveillance camera images on behalf of whom or what, in what situations, and toward what kinds of consequences' (145).

What I conclude from this ethical exploration is that the use of data reaches far beyond the intention of the human agent collecting it (or how the social robot is perceived as), but goes back to the designing of the modules and systems that gather the data, and attached to a future re-use after being collected. This new perspective might be the strongest counterargument I offer to re-address the issues in Chapter One, in which the *good* or *bad* intentions of using social robots as *potential* monitoring devices was debated. I see the undertaken reflection as an ethical antithesis to Chapter One and as an

essential critique of Robot Ethics, a field that operates while lacking these insights on data-related concerns. I further argue that social robots must be taken seriously as potential new information structures, even as dataveillance. Robot ethicists, in general, must highlight that the collection of health data (Knoppers & Thorogood, 2017) from sensitive environments is specifically protection-worthy, since this might be increasingly commodified, as I have mentioned various times.

What I brought forward by aligning the MSS[39] discussions on data to Robot Ethics was that not only must the human intention *prior* to the misuse of a technology be factored into ethical questions and frameworks, but that multiple intentions are already manifested *in* the data inherently; data is ethical from the very beginning on deciding to collect it. As Gitelman (2013) pointed out, *data* is defined with the intention to be gathered by a technology and the concepts that allows its gathering. Therefore, data does not gather or collect itself objectively, neutrally, nor accidently.

Again, the very intention to collect data (and to standardise its collection) is embedded beforehand into the intention and establishment of what kind of data to collect and in what environments, and such data can be appropriated after it is collected. Therefore, data can be collected for one reason, and then be used for another, just as the quality and outcome of derived results can vary according to the context and pattern made from the data (Püschel, 2014: 3). Püschel, Andrejevic, and Laney point to the potential re-appropriation of data for various purposes after it is collected, and this thesis suggests that this use is not limited to the discussion on tracking in the HRI and to gesture or

emotion recognition only[49], since there is no guarantee that this will be the only use for this data.[50]

What has emerged, by questioning intentional and strategic monitoring or surveying, is that the option to survey is embedded in the very structures of contemporary digital technologies. Hence, this allows us to situate the ethical problems early in the exposure or placing of technology. It has not become clear to me how to draw the line between the gathering of data for specific scientific or commercial purposes and the possibility that this very gathering might become ethically problematic dataveillance. An awareness of the possibility for increasing the data sets – in amount and appropriation – is missing as much from Robotics as from the Robot Ethics discussion.

Even if Robot Ethics increasingly addresses the intrusion of privacy, data infringement problems, or the requirement of the robot to collect and process data, the possibility of increased data sets remains unaddressed. The lack to consider this new complexity only exhibits a lack in understanding of data in Robot Ethics and the inability to distinguish between capacities and projections of robots ethically. Considering obtained data gains in commercial value and sensitivity, due to it being gathered in private, yet commercial, health environments (Knoppers & Thorogood, 2017), these potential problems must be taken serious in Robot Ethics. Health care data is, therefore, not only specifically sensitive and protection-worthy, but the collecting of data as a process should be something the elderly patient consents to explicitly, since it is just as much an ethical

[49] The question on the longevity and on the materiality of data should be excluded at this point. It is also acknowledged that tracking or robotic interactions are still very complicated practical processes, which could lead to argument from HR that there is no need to worry about data collection, since, often, robots are not interactive.

[50] I cannot provide additional insights into Data Ethics at this stage (a research area that also moralises technology), but my exploration on data echoes in Floridi & Taddeo's (2016) critique on the disconnection between Robot Ethics and Data Ethics.

concern[51]. Robot Ethics does not take any of these options into account: that monitoring can be problematical beyond a privacy intrusion, and that the developer's intention to use data reaches beyond the context of application or intention.

Data concerns are already troubling social robots and companies. The social robot *Pepper* (a popular and assisting research robot) already faces system breaches and was labelled as an 'insecure' and easy-to-hack technology.[52] Additionally, companies in general – such as with the Facebook scandal[53] – have lost in credibility recently. Therefore, data concerns are not to be taken lightly; neither is a shifting business model that might suddenly have much use of data being available. Just a few years ago, Facebook would have not admitted that their business model is to *not* only connect friends and families, but could be, in fact, to collect and commercialise the data on their users.

Ultimately, Andrejevic is also correct when pleading in his talk, *Towards a Program of Algorithmic Accountability* (2017), that developers and programmers deliberate on how algorithms are made and to what extent these are used. The Facebook case appeared to be publically singling out one person as accountable – the CEO, Zuckerberg – yet many more are involved. However, it showed that there is an expectation that human agents must be held accountable or blamed if technology goes *wrong*.

---

[51] I assume that even the caretakers and human professionals have no idea what the use of robots implies when it comes to data infringement or losing control over this collected data. In the worst case, the roboticists might be naïve on that matter of what the consequences could be.

[52] This article exposes the already existing data and accessibility issues with Pepper, the social robot. Available at: http://www.theregister.co.uk/2018/05/29/softbank_pepper_robot_multiple_basic_security_flaws/ (Accessed 22.01.2019).

[53] More details to Facebook data scandal are available on http://www.bbc.co.uk/news/technology-43649018 (Accessed 22.04.2018) and on https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election (Accessed 23.05.2018).

As the structures of technology are becoming increasingly dominant in forming our information environments, these have also become 'opaque and unaccountable' for human agents to fully control - yet, who else is going to do that? Andrejevic advocates that 'we', the developer/user/university/researcher/journalists, need to create systems which are transparent and allow for more accountability and control. The human agent cannot leave these developments out of sight and unresolved, as they are the only one paying the price. It is important to urge for a multidisciplinary discussion on the status of the social robot as a digital device with an ability to collect data. To ask 'why do we implement social robots in elderly care and who profits from this?' would be a good starting point in making this ethical discussion more critical and of more practical use than current attempts show, which merely moralise the social robot as a perceived human-like agent or a neutral and harmless device.

What I do not suggest is to have more prolific discussions on robots as potentially *bad* agents or *spying* machines without agency, since these debates are flawed on some levels. Instead, I urge for discussions on how the use of social robots is already an ethical process that implies its ethical problems stem from the implementation of technology. Robot Ethics must have a wider data debate as soon and as critically as possible. Even though there are already discussions on this topic, these debate data collection as a neutral process and the ownership of data as a fully controllable decision.

In addition, I encourage MSS to pay much more attention to social robots as well. Momentarily, it seems to me that Robot Ethics is in a trap, since, not fully independent of HR, their discussion might reflect badly on robots as products and make them

unpopular. This is specifically counterproductive to debating new ethics for health care contexts in which more robots will be applied if the statistics are correct.[54]

At the same time, there are already critical debates, even if the field divided robots into superficial categories when discussing their capacities. Another element, however, is that most patient information is being digitalised these days, so, potentially, if the area would be more open on this topic, this could potentially increase the trust from the patients or clients towards robots and the industry. For now, excluding the social aspects of robots from dataveillance debates suggests a negligence and misunderstanding of their ability to collect data and hints to a short-sighted view on digital technologies. In general, I urge for less naivety and much more commercial foresight in scientific and robotic theory, considering robots are commercial products, and, yet, the privacy of a vulnerable group like the elderly must be priceless.

[54] More on the UK Government declaration on the use of social robots in care can be found at https://www.parliament.uk/mps-lords-and-offices/offices/bicameral/post/work-programme/social-sciences/robotics-in-social-care/. (Accessed 12.01.2019)

II. FROM MORAL AGENTHOOD IN ROBOTS TO MORAL AGENCIES IN ALGORITHMS

Part Two summary: Next, I discuss and reflect upon different philosophical positions of moral agency and agenthood in social robots, as discussed in (iii) Machine and Computer Ethics and supported by Humanoid Robotics (HR) and other insights from Philosophy of Technology (POT).

After having contextualised the companion position in HR and the importance of anthropomorphism in this context, I shift to a discussion on computational architecture of robots and the process of tracking, to lay out how to think about such ethically. I then move to morality research in POT to address the ethical questions of computational agencies and their increasing inability to grasp questions on moral accountability. Part Two explores the adaptation of morality into technology – from apparent moral agency, to moral agenthood, moral agencies, moral rules, and moral factors – and the influence of computational autonomy in this process.

I work through three limitations in morality streams and contemporary ethical debates in POT, which I label as; (1) Moral Appearance and Agenthood; (2) Reductionist Morality; and (3) Distributed Morality. The conclusion I draw from these three angles is that agenthood, accountability, and morality become conflated discourses that increasingly remove questions and concerns around moral accountability from applied ethics.

4. On Agency, Autonomy, and Moral Accountability

Chapter summary: In Chapter Four, I continue the discussion from Chapter One, on how to debate moral agenthood in technology. What I illustrated in Chapters Two and Three was that the moral and ethical concerns around social robots in elderly care align to different agency models; either viewing them as companions or as tracking devices. Next, I revisit the theoretical discussions in POT on agenthood, morality, and technology (prefaced by a media-theoretical and Posthumanist angle) to examine the relationship between human agenthood, artefactual agency, and human moral agency from a techno-philosophical point of view. I make use of Johnson & Noorman's (2014) work to guide this exploration, allowing me to comprehend how agency and morality are negotiated within the making and use of artefacts and why accountability is only assigned to human agents.

Before beginning this chapter, I would like to reflect on Chapter One's discussions and suggest continuing with a slightly different discussion on how agenthood is established around the concepts of morality, accountability, and technology, given that the social robot has been scrutinised in Chapters Two and Three as companion, pseudo-agent or monitoring tool, and as a data-collecting device.

Chapter One introduced the idea that moral and ethical questions can differ from each other, but that this is not necessarily the case in the majority of POT literature. While I suggest distinguishing between these to:

1. Highlight how moral concerns anthropomorphise robots superficially and expect them to own capacities or to embed moral reasoning or intentions. This is not a critique on the anthropomorphising of robots in general, as I will discuss later in Chapter Six, but one on drawing ethical standards from this projection.
2. Offer a greater width on agency networks in robots and their linkage to problematic and unaddressed blind spots, such as data collection.

Next, I want to return to a wider theoretical context in thinking moral agency, independent of elderly care, allowing me to explore how morality transforms in POT – from thinking moral agenthood to moral actions, moral decisions, moral norms, moral outcomes, and moral agencies. In most philosophical traditions (moral philosophy and POT equally, but not in Posthumanism), the only agents who can be moral are humans.

In the present context, where social robots are used in professional care, I question what and who is a responsible agent, and if it is possible to hold a non-human agent accountable for its actions. I argue that it is difficult to infer human-like morality from a humanoid body and its actions, and, further, that agenthood becomes increasingly hard

to locate due to the technological complexity of robots. Hence, a dilemma emerges: between the social robot not being human, but being anthropomorphised into one, while the robot owns a huge degree of technological agency and autonomy, which is overlooked.

Arguing that there is a discrepancy between these two agencies is one of the most complex parts of this thesis, since it is this friction of not mistaking the robot as a human-like agent because of its shape, but instead taking it serious enough as technological agency network, that is my focus. This requires a complete rethinking of ethics in robotics, in my view, which might make the humanoid shape of robots less relevant in ethical terms.

Agenthood (used as a singular, stable moment of allocated action in a human entity) is the most important factor when locating agent accountability and can be understood as a relationship between an accountable person and other people, linking performed actions to the accountable agent who is performing them. Noorman (2018) writes in the *Stanford Encyclopaedia of Philosophy* section on *Computing and Moral Responsibility*:

> 'Moral responsibility is about human action and its intentions and consequences (Fisher 1999, Eshleman 2016). Generally speaking a person or a group of people is morally responsible when their voluntary actions have morally significant outcomes that would make it appropriate to blame or praise them. (…) The person or group that performs the action and causes something to happen is often referred to as the *agent'* (Noorman, 2018).

But who or what is 'often referred to as the *agent*'? And why is '*agent'* written in italics? Does this mean that the agent is unknown? That it is still a *person* or that it is not a stable position any longer? These questions become even harder to discuss, considering there is a huge difference between perceiving the social robot as an

anthropomorphised agent, an autonomous computational agent, an instrument, or an ethical dispositif.

I only touch upon two contradictory views on understanding agency of technology, which allow to rethink the morality-led discussions around anthropomorphic projection.[55] One way would be from a media-theoretical angle, which aligns with the Posthumanist view (Pias, 2011; Zielinski, 2013; Parikka, 2015).[56] The second perspective looks at technology as a neutral tool and instrument that, through human intention and implementation of values, can become an extension for human values and morality (Wallach & Allen, 2009). This second way of understanding the agency in technology is prominent in Robot Ethics and Machine Ethics. While I disapprove of this in general, I acknowledge its practical advantages. If the social robot is understood as a media technology, then the instrumental view would not hold up, since technology would not be understood as neutral in this context.

The (material) media view would not allow for any moral accountability within technology, since this question would not even arise in this context. This does not imply

---

[55] 'We have to understand technologies are always involved in a range of what we might term social, political, and technical relations which contribute to any experience of that technology. It is only through an understanding of these relations that we can generate a detailed sense of the nature of a technology, its history and so on' (Kroner & Neyland, 2012: 148).

[56] When referring to a media theoretical view, this thesis follows Pias' (2011) view on what he considers the discipline of *Media Studies*. Pias calls the field less a discipline but rather 'a scholarly interrogation. It is concerned with the question as to how symbols, instruments, institutions and practices contribute to the constitution, circulation, processing and storage of knowledge. In this sense it investigates the media-historical conditions pertaining to knowledge and cognition and therefore is more a kind of historical epistemology. This interrogative approach may not only be found in various academic disciplines; in the sense of »media theory« (*Medientheorie*) it is already to be found in almost every imaginable field of knowledge' (2011: 1).

that the media-theoretical angle understands digital technology as inherently amoral, but media technology is understood as a practice that negotiates moral discourses.[57] This angle does not deny that robots or other artefacts involve a delegated human work force, or are functional to some extent, but it would not support a causal or an instrumental neutrality of technology.[58] Media-theoretical scholars are interested in 'locating the materiality of cultural techniques in technological arrangement' (Parikka, 2015) within technological artefacts and practices. According to this perspective, technological and human agencies are intertwined in one ethical structure, but this view is not represented in HR/Robot Ethics.[59]

There is another way to contextualise technological agenthood, which comes from a Posthumanist acknowledgment of anthropomorphism, which is surprising, considering that Posthumanist ethics is against anthropomorphic and moral views in general. Braidotti (2014) identifies the importance of anthropomorphism in relation to concerns

---

[57] The thesis does not look at moral codes of human conduct, or journalistic practices as ethical exploration of media. In discussions, as Couldry et al. (2013) offer on the ethics in media, they do not take the inherent qualities of technology as already ethical into account. Hence, the human agent is the only morally responsible one in these media discussions.

[58] A different angle on ethics and new technologies comes from a different camp in Media Studies. *Ethics of Media (*Couldry et al., 2013), for instance, does not discuss media technologies as ethical structures, neither as moral ones, but focusses on their use by an ethical human agent. Hence, even Media Ethics, to broaden of the ethical view towards an ethical view on algorithms, would be very beneficial. Another publication worth mentioning in this context is Drushel & German's (2011) *The Ethics of Emerging Media,* in which they discuss ethical challenges for the conduct of journalists online, or in social media practices and the effects online providers have on privacy issues.

[59] This media view on the material and ethical entanglements is not represented in Media Ethics, but partially found in Digital Ethics, in which the complexity of technology is acknowledged, but either robots are not addressed or the technology is not understood as inherently ethical (Couldry et al., 2013; Davisson & Booth, 2016).

around anthropocentrism in a discussion she undertakes with the media researcher Timotheus Vermeulen by saying:

> 'I agree with the distinction Katherine Hayles makes between anthropocentrism and anthropomorphism. You can be a post-humanist and post-anthropocentric thinker. In fact, in advanced capitalism, in which the human species is but one of the marketable species, we are all already post-anthropocentric. But I don't think we can leap out of our anthropomorphism by will. We can't. We always imagine from our own bodies – and why should we, considering that we still live on a planet populated by humanoids who think of themselves as humans, in different ways, with different points of reference? Our very embodiment is a limit, as well as a threshold; our flesh is framed by the morphology of the human body, it is also always already sexed and hence differentiated' (Braidotti, 2014).

Braidotti elaborates on how we (humans) understand agenthood as an anthropomorphic and anthropocentric reference point from which we look at the world (11). According to her, we cannot escape anthropomorphism fully. Chapter Six of this thesis will expand upon this point by tracing the exploitation of anthropomorphism back to HR and the research on social robots.

Both anthropocentrism and anthropomorphism are essential and returning conflicts in my investigation, since these two projective processes represent the major issues I see in how social robots and robots are understood. Social robots are positioned by Robot Ethics either as anthropomorphised pseudo-agents, as robots with neutral tracking modules that could potentially develop the intention to spy, or as agency-less tools misused *immorally* for surveillance. I think this points to the conflict that social robots are exposing in their ethical field; that the moral standpoint is based on the projected

and perceived human-like qualities, while computational qualities of robots are underestimated to the extent that they are seen as ethically neutral and instrumental.

However, Braidotti (2014) picks up the concern of anthropomorphism in a surprising way. She urges us to pay attention to the limitations of both angles – anthropomorphism and anthropocentrism – but she legitimates the anthropomorphic angle as a biological reflex, while she sees the anthropocentric intentions of POT ethics as concerning. While Chapter One already sketched out how Braidotti (2006), in fact, sees morality as being limiting for any ethical exploration (15), she does not ignore the human tendency to anthropomorphise. This indirectly supports my two-fold critical focus thesis; to explore why and how anthropomorphism is exploited and reflected in Robot Ethics, and to also question the anthropocentric moralisation of robots and algorithms (explored in Chapter Seven fully). However, in the wider questions related to accountability and agency, the Posthumanist or ethical media debates are not concerned with asking questions about human or agent accountability within technology.[60]

In the ICT discourses on morality and technology, the agenthood focus on technology, paired with questions on moral agency, is still very much relevant (Kroes & Verbeek, 2014). Here, discussions differ on how much the view of moral agency, computational agency, and computational autonomy are conflated. One concept that is suitable for my exploration is Johnson & Noorman's (2014) overview of artefactual agency that, for

[60] The concerns with accountability are important for Posthumanist thoughts, but not in applied and practical questions on when does a robot behave badly or intend to harm someone.

them, is still clearly detached from a human moral agency and accountability. In their framework, a robot could be seen as either an artefact or an instrument.

I consider both angles as insufficient to understand robot technology, but worth picking up to make use of certain valuable insights, even only as a reference point to reflect upon. For instance, I deny the instrumental view of technology (in general) as a theoretically holistic way to understand technology, since I reason (as supported by media theorists and Posthumanists) that a deterministic and instrumental view on technology does not consider the growing autonomy and complex agencies intertwining. However, as I have already highlighted many times when mentioning what a single moral agenthood allows for, thinking of technology as a tool offers the same huge advantage; it makes the allocation of accountability much easier. Gunkel (2016) writes that:

> '[T]he instrumental theory of technology not only sounds reasonable, but also is obviously useful. It is, one might say, instrumental for figuring out questions of moral conduct and social responsibility in the age of increasingly complex technological systems. And it has a distinct advantage in that it locates accountability in a widely accepted and seemingly intuitive subject position: in human decision-making and action' (238).

The instrumental view cannot hold up to the increasing complexity of digital structures, since it has 'significant theoretical and practical limitations, especially as it applies (or not) to recent technological innovations' (237). It might be interesting to pick up what Clifford Christians states about the instrumentalist view of technology as being linked

to how descriptive ethics are understood in industrial countries. According to Christians (2011):

> '[T]he focus here is on instrumentalism as the major challenge for doing descriptive ethics in technologically sophisticated countries. The prevailing worldview in industrial societies is instrumentalism – the view that technology is neutral and does not condition our thinking and social organization' (16).

Supporting an instrument view on technology would offer a simple answer on accountability in the debate. This would allow us to quickly resolve the accountability dilemma related to social robots as tracking devices. But, in order to do this, it would also require us to reduce the complexity of social robots to neutral tools without much computational agency, and to deny their companion position equally, since, as neutral tools, these cannot be mistaken for humanoid friends or agents either. Accountability would be much easier to locate in the human agent (only) using a technology (assuming social robots are a technology), but this also suggests neglecting what Johnson and Noorman see as the artefact's agency and computational autonomy (2014: 144).

Johnson & Noorman's approach suggests distinguishing between what kind of agency the artefacts always own and what kind of *additional* agency they are given by humans to extend a human action. They are concerned with how an artefactual agency connects to the exhibition of a moral autonomy.[61]

---

[61] Johnson & Noorman (2014) do not differentiate between artefact or technology. This becomes clear when they include the examples of artefactual agency being in a CCTV camera or in software agents. Therefore, it cannot be assumed that they see the artefact as different in this debate than, for instance, computational systems.

This approach relates to the recent discussions on digital materialism (Pötzsch, 2017) advocating for a distributive model of agency in media theoretical and Posthumanist work. The difference would be that the material and Posthumanist discussion would attribute inherent agency to every artefact, and would not agree with what Johnson & Noorman (2014) are about to argue, namely that technological agency is an extension of the human agency. For Johnson & Noorman, the agency within artefacts, as man-made devices rather than stones, for instance, can be understood in three different steps in which agency moves increasingly towards moral autonomy (without reaching it). First, the causal efficiency; second, the 'acting for' agency; and third, the moral autonomy (Johnson & Noorman, 2014: 144).

The causal efficiency perspective relates to the Actor-Network-Theory (ANT). In this theory 'the causal efficacy of each node in a network is dependent on each other' (149), whereby the nodes refer to how humans, technology, and environment co-shape each other. This view does not favour an anthropocentric agent position, but equally does not place any moral autonomy in artefacts. In the second step, the 'acting for' agency (149), the human agent decisively uses an artefact with an intention to achieve something or to replace another human agent through its use. This taps into the problem in which, for example, the robot as a caretaker could be bound up, because, on the one hand, these robotic devices are supposed to act as professional caretakers and fulfil even pragmatic jobs like monitoring medication or safety, but, on the other, they are supposed to be perceived as companions, as human-like agents (see Part One, Chapter One).

Johnson & Noorman argue that Latour's work would support this suggestion in combination with the first position. The tasks of machines are hereby understood as a delegated human 'program of action' (Johnson & Noorman, 2014: 149); for example, in the case of automatic doors that open the door as a substitute for a human agent. But automatic doors have no moral autonomy and neither are they expected to have any. The important part here is that the door is given an 'acting for'; as in, a *metaphorical* position for human agency (150). They write: 'Here agency involves representation, though the representation involves the agent using his or her expertise to perform tasks for the client' (149). This position is built on an agreement of the difference between agency and accountability. The automatic door has agency (and, to an extent, autonomy), so it can also have consequences when hurting someone, but at no point is it an agent, neither can it be held morally accountable due to its having a *bad* intention to hurt or trap someone by malfunctioning.

The theory of the 'acting for' agency extends beyond *the* visible context of delegating perceived actions. Johnson & Noorman bring up the example of 'software agents' as an extension of human understanding of code (150). I do not agree with this statement, since I consider that it conflates defining an agent only metaphorically, and having an accountable agent to refer to. This view suggests seeing software as an extension of human tasks and for this process to be understood as neutral or causal in its unfolding. It is not possible, however, to identify a linear extension between human and software, despite software being linked to human language and semantic codes.

Also, this theory allows for every possible technology or artefact to be associated to a human origin and oriented towards it. This is a highly anthropocentric and problematic way to understand non-human entities – as robots are. Further, thinking about a software program *as if* it has a single agenthood is symbolic, but not practical. Would the data, the algorithm, one bit, or the program language be this agent? Most importantly, pinpointing software as an agent sounds impossible, since one is fluid and the other stable. However, Floridi (2014), who will be revisited in Chapter Seven, offers a similar concept on algorithms as 'mindless agents'.

According to Johnson & Noorman (2014), there is an important distinction worth making. Machines, as much as software, are understood as delegates for humans that substitute for human actors by drawing the attention to the role they perform in shaping human actions and morality. These are not understood to be *actual* agents, but only *metaphorical* agents. It is tempting to take complex technologies such as robots into this view since they *look* like humans, but this is a simple association that HR makes use of and encourages.

I suggest a thought experiment at this point to clarify why this theory is included in my exploration, despite the problems of being simplifying and instrumental. What I advocate is to try and see social robots as if they were agency extensions of a human. This could, as I believe, reopen a new perspective on social robots and also on tracking; one that crystallises why the social robot is only a visual metaphor for a human body,

and ignores that to be human means more than just appearance – it could also refer to judgement, aliveness, and movement.

In defining the social robot as a *human-like* machine, we are pulled up on a crucial point: what is the human *program* (as in ability) that robots are supposed to be 'acting for'? In other words, what element of the human agency is it that they are *extending* (assuming that they extend human agency, which is a simplistic view, but that should be overlooked in this example)? Following Johnson & Noorman (2014), the social robot would contain a human *program* that mainly refers to a visual similarity to the human agent. It has the same body shape even if it is not similar enough to be mistaken for an actual human being. At least, this is the metaphor that HR is going for, as I will show in Chapter Six. This *metaphor* has consequences on various scales, since, if agreeing on it, then the social robot is designed to suggest some sort of humanness, aliveness, or intention, but cannot deliver on any human qualities beyond a human visual alikeness and a perceived agenthood.

There is another side to this, however. The robot is in some ways *alive* or *moving*. Not organically, but technically. It does perform actions and processes; something is moving in its shell, to make the outside body move too. For instance, if I return to tracking and think it through the metaphor analogy, tracking could be seen as a metaphor for interaction or for surveillance equally. I argue in this thesis that these aspects reflect on the epistemological framework from which the definition of programs transfer into the ethical discussions. This means tracking can be metaphor loaded to symbolise

interaction and sociability as HR sees it. Or, if we go with the insights from MSS and data infringement, tracking can become a metaphor for dataveillance. In fact, it is both simultaneously, since intentions and action overlap in the plurality of data. The point is this: What we consider the robot to extend might not be human alikeness, but an equally important co-agency built on interactivity, one that must be thought beyond extension.[62]

Another metaphor I want to emphasise is related to the previous one, but amplifies the ambitions that HR and Robot Ethics have for social robots. For elderly care, the social robot is a caring technology (not exclusively, but often). This can mislead the human interacting with robots, since the view on social robots as *caring* machines is purely drawn from the similarity of them performing human gestures and expressions (the theatrical play, *Spillikin: A Love Story*, that I mentioned in the introduction is a good example for this).

However, what needs to be considered is that the metaphor of a technology and the actual functionality might not overlap. This way of thinking technology is also very anthropocentric. It can even be argued that the companionship metaphor does mask the robotic qualities with the human similitude of the social robot. Here, encouraging anthropomorphism creates a problematic analogy in which the metaphorical use of a technology overrides its actual computational ontology. Hence, to position social robots strategically as perceived companions and as *caring* technology means that the elderly should expect these robots to be empathetic, to engage, to listen, to be kind and

[62] Tracking and robots are also always an *extension* of human agencies, since these are made and programmed by human developers.

attentive. These expectations complicate the discussion on moral agenthood hugely, because these are built on a conflation between perceived metaphorical and invisible ontological qualities. Furthermore, because these lines are blurred, the discussion about what social robots are 'acting for' gets increasingly complicated.

The importance of clarity and precision on the metaphors and their encouragement is crucial for various purposes. Not only do I emphasise this, but Johnson & Noorman (2014) also warn that these metaphors are not innocent concepts; they can sometimes even be dangerous. They draw attention to particular similarities between two things; using one that is presumably well understood to help understand the other, that is not. However, in thinking of robots metaphorically, we may be directed to think that the two things have more in common than they do (150, 151). I think this level is crucially problematic, as the aims of HR are feeding into Robot Ethics, allowing for anthropomorphism to encourage the position of social robots as human-like, or as companions, or as perceived agents in ethical discussions, but denying the invisible ethical level to be addressed.

I argue that the first angle is insofar a limited, limiting, and *wrong* metaphor to draw. What I mean by *wrong* (a term Johnson & Noorman use) is not referring to it being a deceptive error, and it must not be mistaken with Turkle's critique (2011) in Chapter Two. Instead, I am concerned with the deceptive element when drawing ethical accountability from the superficial alikeness, or to encourage the human agent to mistake the robot as actually having human similitude. Thinking that the social robot is

somehow *human-like* on the grounds of its appearance might be an aesthetically seductive fallacy only, but it leads to the potential neglect of the robot's (in my view) more important capacities as a non-human artefact. This is what Johnson & Noorman warn of when relying on representational similarities, namely, to ignore the 'important and relevant dissimilarities between the compared entities', which 'end up being pushed to the background by making a particular analogy between the two entities' (151).

Social robots cannot be fully understood with this model, however, though it was helpful to me in illustrating the metaphorical dilemma, which is found in the perceptive construction in agenthood and the companion position the robot is given in elderly care. However, this did not allow me to clarify any accountability questions, since the artefact is not given any accountability by Johnson & Noorman. The 'acting for' theory acknowledges that artefacts can always have ethical consequences beyond the questions of being extended human programs. The example of the mechanical door briefly addressed how a malfunction could hurt someone but not be held accountable for it. Hence, the 'acting for' model comes with ethical consequences even if not moral accountability (152).

What Johnson & Noorman have outlined by discussing the difference between agency and moral autonomy is reconsidered and challenged by Wallach & Allen (2009) and Floridi (2014) in Chapter Seven, when the debate on computational autonomy of technologies is aligned with morality. Johnson & Noorman (2014) conclude, in terms of the different levels of artefactual agency, that none of them relates to the moral

accountability or autonomy in human agents. It is important for them to acknowledge that artefacts always have agency and that 'artefacts make a moral difference' (152), but these do not get elevated into owning moral accountability. They are clear in their opinions that accountability and responsibility are missing in artefacts; hence, they do not go beyond the 'acting for' step in their trifold agency concept. The third step in their model is not outlined in detail here, since it is the simplest one. For them, moral accountability is always assigned to the human agent as the only morally accountable agent independent of how much agency the artefact has been given.

I address the confusion between 'acting for' and moral autonomy as a huge challenge within (i) Robot Ethics and in (iii) Machine and Computer Ethics. The reason could be that the metaphorical definition of a technology is not suitable to grasp the ethical dimension, since it follows an association chain that, again, is anthropocentrically thought and encouraged, but with no justice in understanding how certain technology operates. What I think that Robot Ethics and Machine and Computer Ethics do is to correlate computational autonomy with computational *moral* autonomy, because it is argued that human moral agency is linked to human moral autonomy (151). Therefore, computational autonomy could be something like a *computational morality*. This inference is a problem in my view, since even if these concepts all relate to each other, their implications do not derive from the same ethical discussions. Computational autonomy, for instance, can derive from a mathematical accuracy built on the ability to follow a set of rules. Still, the question about having an awareness of the set of rules or

their making is not sufficiently considered (debating *consciousness* is not discussed here).

Johnson & Noorman (2014) point to the issues with debating *autonomy* in both entities. Their view on artefactual autonomy corresponds to an 'independence of things from immediate control by humans' (152). To an extent, this is what Broadbent (2017) asks robot autonomy to be able to do.[63] The next chapter returns to the initial critique that begins Chapter Two, of how the social robot is only perceived as an agent, which supports a position I (and Coeckelbergh) argue to be the first limitation I draw on in (1) Apparent Morality and Agenthood, which I will discuss next.

[63] Christians supports what Johnson & Noorman (2014) argue as a dangerous step: to view technological autonomy and technological moral agency as correlating qualities. He writes: 'Autonomous moral agents are presumed to apply rules consistently and self-consciously to every choice. Through rational processes, basic rules of morality are created that everyone is obliged to follow and against which all actions can be evaluated. In communication ethics, neutral principles operate by the conventions of impartiality and formality. This is an ethics of moral reasoning that arranges principles in hierarchical fashion and rigorously follows logic in coming to conclusions' (Christians, 2011: 2).

5. On Moral Appearance, Agenthood, and Moral Philosophy

Chapter summary: In this chapter, I survey wider robot agency models and make use of an existing critique from POT to also reflect on how agency and morality models from moral philosophy are applied simplified in POT. My major focus will be on the (1) Moral Appearance and Agenthood model to show that judging human morality as exhibited in action or appearance cannot be simply adapted to robots. I already offered insights on why that view is limited; now, I also expand on inner-disciplinary concerns in POT. As I brought up already, deriving moral agency from the humanoid shape of the robot and its visible actions limits the full ethical discussion in Chapter Two to seeing data concerns clearly, but it further supports what Coeckelbergh calls a 'psychopathic robot' (2010) model. I conclude this chapter with a study by Kahn et al. (2012) to highlight the dangers of superficially assigned morality, as this feeds back into drawing any moral accountability of robots from their appearance by projecting.

I focus next on the relation between the making of robot agenthood and the appropriation of traditional moral philosophical streams, which are either simplified or reduced to allow for a technological model of morality to emerge. I hereby return to social robots again, but not to elderly care. As argued various times, looking at social robots as companions only reduces their agency model to a perceived one. However, I identify that, within POT streams that leave Robot Ethics and elderly care, more critical voices can be found on what I label as the 1) Apparent Morality and Agenthood model, which conflates between the qualities that are apparently embedded in social robots and those which are owned. What becomes clear to me is that as traditional moral philosophical streams are being adapted to POT, so are their old concerns taken over as well; for instance, thinking in dichotomous pairs between appearance/essence of robot morality or intelligence.

The robot ethicist and machine morality researcher Wendell Wallach (2010) brings to our attention the historic origins of the study of robot and machine morality. For Wallach, the different routes of morality concepts lead to the complexity and inconsistency of the field, and, as I remarked as well, the perspective through which morality is discussed creates different ethics of robots. It is not surprising that Wallach writes:

> 'The study of moral decision making is profoundly influenced by a long-standing tension between moral philosophy and moral psychology. Moral philosophy and moral psychology developed hand-in-hand. However, throughout the 20th century two philosophical positions, the is-ought distinction and the ''naturalistic fallacy'', served to buttress a division between these fields of study' (244).

What this quote further illustrates is not only a divide between ethics and morality, but between the conflicting perspectives in the philosophical or psychological paradigms on morality. According to Wallach (2010), this conflict is mirrored in the conflict between moral sentiments (as found in Hume's and Smith's work) and moral norms, as a gap between 'ought' morality and 'is' (244) morality, but also between values and actions. He explains that:

> '[T]he is-ought distinction is broadly understood as a fundamental gap between all descriptive or factual statements and normative or prescriptive judgments. This can mean that understanding the psychology of how people make moral decisions does not inform us about what people ought to do' (Wallach, 2010: 244).

This connection between the psychology of moral actions and the inference to moral reasoning[64] is crucial for robots to be understood as morally competent. Malle & Scheutz (2015) expand this to an inference between actions and values by asking: 'What *would* it take for robots to be seen as morally competent?' (2015: 486). They hereby highlight that the *apparently* good actions are sufficient for robots to be seen as good agents. In addition, POT streams on robots rarely point out specifically what moral philosophical school they assign their views to (Wallach is an exception). Hereby, Kantian, Utilitarian, and virtue ethics are popular backdrops that emerge often in the discussion on robotic morality and agency as entangled concepts, since some POT literature is not always clear in what traditional morality concept is translated into the

[64] I argue that Luhmann's critique in Chapter One is, in fact, a critique on the psychology of morality and on the pathological view behind moral reasoning as a deficit oriented system without clear answers on what makes humans morally good or bad.

techno-morality discussion. However, these three traditional ethical models are the most valuable for POT and Robot Ethics for several reasons. Each of them is, in some way, either anthropocentric or anthropomorphic, but most are built on a single agenthood idea (Floridi and Brey will be exceptions)[65]. To recall, the anthropocentric agenthood (looks at structural and inherent capacities) is discussed later in Chapter Seven, whereas, for now, I am reflecting on the anthropomorphic agenthood (looks at perceptive and apparent qualities or actions).

Briefly outlined, the Kantian view on morality is appropriated to create a mechanistic, rationalist action and rules-based robotic morality as aligned with the idea of duty[66]. As Christians (2011) argues, to focus on Kantian ethics allows us to infer robot moral agency from robot action and appearance on the grounds of being able to draw moral reason from visible actions in humans.

> 'For Kant, reason demands moral action. It is the nature of reason to will universal law, and it demands this not only in theories of science, but in practical thinking about what we do. Hence, we ought to base morality on reason. Reason is my authority for acting morally' (3).

While the 'utilitarian perspective applies the principle of utility to individual moral actions and the rule utilitarian applies the principle of utility to moral rules. The right

---

[65] Dumouchel & Damiano (2017) raise issues with a single agenthood view as well. However, they join a robot ethical research from a crucial perspective, trying to establish a new ethics around social robots and anthropomorphism, but it is one I struggle with, due to its perception focus. However, they critically endorse that it would be a mistake to position the robot as an individual, or as a judging agent, instead of seeing it as 'moments of a complex technological system' (191).

[66] 'Dutifulness reflects good will and the desire to do things right based upon rules that everyone ought to follow. That is, a dutiful person acts the way they do because of a moral rule. These rules are imperatives that are either hypothetical or categorical and they are the means by which reason commands our will and our actions' (Herschel & Miori, 2017: 33).

act is one that produces the greatest happiness for a community or society' (Herschel & Miori, 2017: 33). What I consider a problem is that, within various research papers on these discussions, the terminology around *good* and *bad* virtues is not questioned but considered as universal, which is also heavily critiqued by the Posthumanist Braidotti (2006). These two ethical perspectives, Kantian and Utilitarian ethics, further differ theoretically from Virtue ethics[67], but both do support an alignment of actions to *good* virtues while encouraging an anthropomorphising of robots.

This linkage between an *apparent* agenthood and qualities of the agent occupies wider robotics research as much as another theoretical camp that orients itself often around the dichotomy between perception/essence: Artificial Intelligence (AI) research.[68] In my view, the conflict between 'weak AI' and 'strong AI' (Duffy, 2003) reinforces the distinction between apparent *as if* qualities and actual qualities of digital systems and robots[69] even further. I consider this distinction in research as conceptually problematic, on the grounds of allowing for rhetorical imprecisions and the assumption that understanding robots or computers allows for any perception-based view on their qualities. The problem being that if the human perception relies on *as if* intelligence or

---

[67] 'Virtue ethics' consideration of individual character is undoubtedly an important starting point for developing ethical motivation and awareness but (…) the lack of a rational and systematic analysis might leave any number of publics or alternatives unconsidered in the decision' (Tilley, 2011: 197).

[68] It is difficult for this thesis to distinguish between 'weak' and 'strong' AI (Duffy, 2003), since the intelligence of a robot is not considered to be visible or perceptive; it can also be computational and non-human. Furthermore, Damiano & Dumouchel (2017) have pointed out that the AI research canons do not fully agree on what AI is (x).

[69] This conflict goes back to Turing's research (1950) from which he concluded that there is no actual artificial intelligence. In this example, computer systems are only *as if* intelligent, because they can only follow the rules, but not think autonomously (Johansson, 2011: 9). Dumouchel & Damiano (2017) challenge this dichotomy with their work on Internal Robotics.

as if morality, (encouraged by keeping the *good* virtues and the ability to act *good*), morality and intelligence are reduced to agenthood deprived concepts and are bound to what *appears* to be moral or intelligent reasoning. This reflects into the ethical theories on robotic morality that focus on a perceptive view of morality (and emotions) (Malle & Scheutz, 2015), as I exposed.

The *perceived* or apparent agenthood position of robots is not simply accepted without certain critique within the fields researching it. The philosopher Mark Coeckelbergh (2010) warns of designing 'psychopathic' robots (235), which can execute perceivably morally correct actions and tasks by themselves (to a degree, autonomously), but still lack any awareness of what moral value is, which encourages the perception that they are detached from their technological capacities. Coeckelbergh specifically highlights that the ability of emotionality must also be linked to the ability of morality. The fact that both concepts are understood as perception-based and expression-based in robotics only support, for him, a highly problematic inference that keeps the moral qualities of the robot aligned to the perception of the human engaging with it.

To build moral agency as aligned to visible action-based performance, which relies on an association chain between actions and moral values, is concerning to me, since robots are assigned agency without consciousness and this happens independent of what moral school has been chosen (differentiating between a Kantian or a Utilitarian model is hereby irrelevant). Hence, this morality model, as Coeckelbergh and I agree, does not account for the robot being able to reflect on the actions it performs. What moral

appearance encourages is a problematic and intertwined causal link between appearance, actions, and robot virtues to ground onto the *perceptive* moral agenthood of social robots. Coeckelbergh (2010) states:

> 'Consider discussions of psychopathy, it is suggested that psychopaths can follow rules but do not have the capacity to feel that something is morally wrong. Even though they may act in accordance with moral conventions' (235).

His research (2009, 2010) explores how the human-like appearance reflects problematically into morality and *apparent* morality. For him especially, the Kantian approach leads to new limitations (which also come from being appropriated instrumentally and simplified), because it does not consider the influence of emotions and imagination as, for instance, Humean ethics would (Coeckelbergh, 2010: 235).[70] Further, the human-like appearance of robot actions cannot be related to a moral responsibility. He writes:

> '[W]e (will) interact with humanoid robots as if they are human. We need not know their 'mental states' for blaming them, for treating them as companions, or even for loving them. Both the ascription of agency and of responsibility are, in practice, independent of the real[ity]. I coin the terms virtual agency and virtual responsibility to refer to the responsibility humans ascribe to each other and to (some) non-humans on the basis of how the other is experienced and appears to them' (2009: 184).

Even if moral agenthood cannot be derived from appearance ontologically,

---

[70] Coeckelbergh further argues that 'if we want to build moral robots, they will have to be robots with emotions. But can such robots be built? In what follows, I first argue that it is not likely that in the foreseeable future we will be able to build such robots, because to do so these robots would have to be conscious, they would have to have mental states, and we would have to be able to prove that they have these things' (Coeckelbergh, 2010: 23).

Coeckelbergh (2010) does not abandon this possibility fully. He instead suggests, in my view, an unpractical middle ground as a solution: to understand apparent morality of a robot only as being 'virtual'. Even if he clearly argues why robotic morality cannot be simply a simulation of Kantian ethics and an application of action-led rules, he allows for a virtual model to substitute this view. He maintains that humans would not make a robot more accountable than a computer, and yet views them with a certain degree of agency nonetheless.

Coeckelbergh (2009) develops the idea of a 'virtual moral responsibility' (183) that computers and robots can be assigned to, which would suggest a less accountable version than expecting their actual moral responsibility; a model I consider as a *light* version of morality. He advocates this model from the perspective of how we identify moral agency in human agents. Since, we (humans) never *really* know if other human agents have or commit to an internal moral responsibility or value system, we ascribe moral values to our actions and appearance. Hence, by assigning some form of morality to what we perceive as a moral action is normal. For him, every form of moral responsibility is, more or less, always *virtual*. While this might be true, I ask if he underestimates the idea of trust and the fact that we have learned to know what to expect from a computer, but not what to fully expect from a robot, especially if it smiles at us. Still, for him, since this is how we perceive computer or technology agency as well, it is a legitimate approach to understand robots. He states:

> ‘A different way of applying the label ‘moral’ to things is not to focus on the transfer of intention and responsibility from subjects (humans) to objects (artefacts), but on the transfer of value. We humans not only design artefacts, but also give value to objects, artefacts, and other non-humans’ (183).

I have huge reservations with this ‘virtual’ or *light* morality version, since it is highly anthropocentric and leaves the question open: with what is he concerned? With understanding robotic qualities and moral reasoning? Or with the distinction between moral actions as outcome-oriented or morality agency as built on agenthood? It remains unclear for me how the ‘psychopathic’ robot, that he initially critiqued as having no awareness of its actions, is transformed into a *virtually* moral agent to avoid inferring moral agenthood into its agent-like appearance. While theoretically interesting, I am not sure about the practical value of this approach, since I do not know how humans should identify the difference between apparent, virtual, and actual responsibility.

Coeckelbergh’s (2009) own conclusion urges for a ‘proactive ethics that intervenes with its evaluation at the design stage rather than when the artefact is used, [so] we [can] better think about the ethical problems now’ (Coeckelbergh, 2009: 186). The hindrance in this discussion is, for me, still in the overestimation of the *perception* of humanoid robots when it comes to their ethical agency. Especially, since their design is kept specifically anthropomorphic to evoke an agenthood these devices do not have, what does this suggest to the human interacting with them? This allows, in my view, for a twofold mistake; that the robot owns *agenthood*, and that it owns a *moral* agenthood. This also leads to confusion about what kind of concept is taken into consideration when understanding moral agency; is it a duty, action, or is it virtue-based?

I hopefully provided, with Coeckelbergh's support, more clarity on why morality and agenthood cannot be grounded in appearance, since accountability and responsibility – being important factors – drop out or must be aligned *virtually* and in a lighter version to then raise more confusion.

I claim that agenthood can only be situated in a human subject (though, I doubt the singularity of such as being a linear or stable reflection point or entity), who can be held accountable for the actions they perform. This aspect is one major advantage in holding onto anthropocentric ethics; otherwise, the Posthumanist view on technologically fluid and intertwined agencies could be more appropriate to address the ethical problems around Robot Ethics. Problematically, neither Robot Ethics nor HR is interested in Posthumanist ethics, since both fields consider a model of human agenthood and morality as their core upon which social robots are built. However, on the other hand, if the modelling of agenthood is exclusively perception-based, my worry is – as supported by Malle & Scheutz (2015) saying that moral 'decision-making' in robots cannot go beyond a 'quasi-moral' process (487) - that this produces 'psychopathic' robots only as Coeckelbergh correctly assessed.

I identify another issue emerging, which is aligned to the previous one: Not only can agenthood be aligned to a perception and appearance model of robot morality, but so can autonomy. Scheutz (2012) demonstrates this with a different path that further complicates the discussion. He points out that not only can moral agenthood be perceptive, but that autonomy can be as well. For Scheutz, a *perceived* autonomy is

essential to position the robot as a 'social agent' (2012: 216). This statement, in my perspective, only creates an oxymoron between having an autonomy but without agency, or vice versa – a conflict he does not resolve, by writing that 'perceived autonomy is so critical because it implies capabilities for self-governed movement, understanding, and decision-making' (216). How would non-perceived, computational autonomy be judged?

What I find problematic in his quote is that autonomy is, again, caught up between a 'perceptive' but also a 'critical' (216) capacity, which must not be aligned at all. It is not only Scheutz's work that demonstrates conceptually confusing arguments; these are continued into morality research as outlined through the following research study that further exhibits what I label the limitations of thinking an (1) Apparent Morality and Agenthood.

I want to conclude this chapter with a critical reference to research by Kahn et al. (2012), which drew my attention regarding their legitimising of an apparent morality in an experiment. In this study, the researchers made participants decide on moral accountability of a humanoid robot named *Robovie*[71], based on their impressions of its actions and looks. The experiment that Kahn et al. undertook is on how much moral agency or awareness can be drawn from the robot's behaviour and from the interaction with an interviewer, and on how moral a robot *appears* to be. For me, as a non-roboticist, the information and outline given in the research paper on the experimental

[71] Kahn et al. (2012) use the term *humanoid robots* synonymously with social robots.

set-up was confusingly loose, in the sense that it remains unclear if the robot is meant to execute a pre-programmed and scripted dialogue and set of movements, or if it is supposed to be responsive to outside input (for instance, making use of its ability to track to the environment).

This point is not irrelevant. In fact, it makes a difference if the study's aim is to explore the computational autonomy of robots as being morally aligned, or if it set out to explore what visual clues the human participants require to be able to anthropomorphise a non-responsive robot, so as to understand it as a morally accountable agent (even if this robot is none in both cases). If I interpreted the experimental set-up correctly, then the use of two 'integrated interaction' (24) patterns within the robot's programming indicates the use of a scripted and pre-recorded HRI scheme, and implies that the robot is not responsive to the input, but is remote-controlled, while the participants might not be aware of this lack.[72]

In the experiment, the robot performs/acts out a dialogue (the so-called *interaction*) with a human interviewer who is meant to involve the robot in 'morally challenging' topics and conflicts (24). At this point, if accountability is highlighted, it would be already ontologically impossible to account any moral agenthood to the robot, since this robot becomes exemplary for the 'acting for' model, when recalling Johnson & Noorman (2014), and would be nothing but the extension of human agency. In the experiment, the human participants are required to judge whether this interaction is linked to moral

[72] This might be what Scheutz meant previously with 'perceived autonomy', which already troubled me then as oxymoronic ontology and as problematic. It does so again.

accountability, but what the researchers do not reflect on is that the perceived moral agenthood is detached completely from any agenthood and partially from acting autonomously (maybe not from a computational one, but this must be explained later).

Therefore, the experiment is flawed in my view, but it exhibits the tendencies I keep bringing up: the superficial implications of good behaviour aligned with actual moral agency, robot technology agency, and anthropomorphised agenthood. These terms get conflated throughout. I think this kind of research does not say anything about the robot as a technology or as a potential moral agent, since this evaluation of morality remains an associated and superficial concept only. It almost appears as if the motivations for this experiment were to test human psychology, and are on the human abilities to anthropomorphise robots, rather than being interested in the establishing of robot moral agency, which, for most roboticists, is a rhetorical buzz word after all.

The researchers' conclusion only confirms what I believe robotics often relies on: the human inability to distinguish robot perception from robot capacities.[73] This shows in the ranking of several associative categories with which the participants had to either agree or disagree. The participants are supposed to answer on, for instance, the robot appearing to have mental states (73 per cent agreed), or feelings (35 per cent agreed), or being a social other, a friend (70 per cent agreed), or being conscious (50 per cent agreed) (37). 65 per cent of participants answered that *Robovie appeared* to be morally

---

[73] This might be a concern in human psychology as well, since we can only read cues and expressions and not each other's minds, but I do not support how robots are aligned to being in the same psychological framework.

accountable (the researchers correlate accountability of robots to that of a vending machine, which is on the other end of the morality scale in this scenario). Kahn et al. further conclude that because of these results (which seem predictable for them), humanoid social robots must be held at least '*partly* accountable' in terms of their moral status (39). From this, they advocate for a new ontological hypothesis allowing for 'personified robots' – as much as other 'embodied personified computational systems' (39) – to be defined on the grounds of these technologies becoming more pervasive.

I strongly disagree with what Kahn et al. (2012) advocate for, since they have created a problematic ontology[74] by aligning evaluative moral capacities and ability to judge to perception and to anthropomorphism. It simply cannot be enough to believe robots look *nice* if they do not understand niceness. However, this idea initially snuck into scientific thoughts and research, so it shows that the ethical or moral discussion is not taken seriously enough, since ethics, widely, and the consequences of unethical behaviour are not. *Do People Hold a Humanoid Robot Morally Accountable for the Harm It Causes?* – their paper's title – is not only an innocent question but has severe consequences if the answer is 'yes'. Conflation and confusion should not be encouraged but, conversely, the necessary distinctions must be made as precise as possible, stating that non-experts

[74] More on how to understand ontology in computational systems can be found in Man's (2013) work *Ontologies in Computer Science*. She writes: 'Research on ontology is becoming increasingly widespread in the computer science community, and its importance is being recognized in a multiplicity of research fields and application areas, including knowledge engineering, database design and integration, information retrieval and extraction' (43). My major point is that computational qualities and programming capacities have to be considered when discussing the computational ontology of the robot's *being*, not only of the robot's *appearing*.

cannot guess these. Further, I doubt what value can be taken from a perceived moral agenthood for the ontological understanding of technological complexities. To align appearance with qualities or non-human systems with human organisms remains a very confusing approach towards progressive technology agency models, in my view.

Hence, the unresolved question for me remains: What does the appearance of a computer/robot say about its moral or agent abilities (assuming it could develop some)? This experiment confirmed the perceptive dominance of the robot's design, one that is detached from any understanding of computational complexities. I do not see how the participants' anthropomorphic response could confirm what Kahn's et al. (2012) suggest becoming a new ontology, but I notice the opposite happening; it confirms that no new ontology is established. While, in theory, they are correct about robots requiring a new ontological approach, their reasoning does not mirror what they in fact explored in their experiment.

Rightly, Dumouchel & Damiano (2017) have suggested that we should rethink robots as a new 'social species' (xiii) to overcome the dichotomy between how a robot appears to be and what a robot can do as a computational system. As I have pointed out various times, the agency models influence ethical concerns; therefore, a rethinking of robot agency in POT and Robot Ethics is urgently needed. I will deepen this aspect in Chapter Six as I provide insights on how this humanoid perception fallacy has evolved around social robots and why it is still important and central in HR and Robot Ethics.

In the next chapter, I leave the ethical canon behind for a while by entering HR

research, in which ethical issues around social robots play a minimal role. Chapter Six is an extensive and contextualising chapter on the background of the *companion* robot agency and on the value of anthropomorphism. It stands out as a chapter, since it looks at the social robot historically and epistemologically, by focussing on the importance of the humanoid robot design, the social context of using social robots, anthropomorphism as a strategy to increase the robot's sociability, and the importance of computational interaction and tracking modules. I use it to outline how the social robot is, in fact, two things all along: the companion robot and the tracking device. This returns to the initial conflict I presented in the introduction.

6. Understanding Social Robots in Humanoid Robotics

Chapter summary: I extensively contextualise and analyse the social robot as a companion in this chapter to point out that this position aims for a paradigm shift in how we perceive robots: from being only machines to becoming human companions. At this point, I have raised various concerns on the companion agency model, but now, I am explaining its origins and its purpose. First, I survey the literature on social robots within Humanoid Robotics (HR)[75] to provide an overview on social robots in HR through research from pioneers such as Breazeal (2002, 2003), Lee et al. (2005), and Fong et al. (2002). The first section focusses on the importance of anthropomorphism as a projective induction using Duffy's (2003) and Lemaignan's et al. (2014) work. I then move to the importance of computational sociability and of the tracking ability, which I examine as an ethical process by reflecting on the FACS model by Ekman. I conclude with two interlude discussions on, firstly, a new way of thinking about anthropomorphism ethically (Damiano & Dumouchel, 2018), and, secondly, on the rhetorical concern in HR and AI, which intersect with robot agency discussions.

[75] Humanoid Robotics (HR), a conglomerate of research from Robotics, Psychology, Mechanics, Engineering, and Computer Science, explores this social robot by: '(…) studying, developing and realizing these socially capable machines, equipping them with a very rich variety of capabilities that allow them to interact with people in natural and intuitive ways, ranging from the use of natural language, body language and facial gestures, to more unique ways such as expression through colours and abstract sounds' (Read, 2014: i).

The Social Robot – From Machine to Social Companion

Summary: The first section in this chapter summarises an important paradigm shift in HR; from positioning robots as machines, to designing social robots that are perceived as human companions. I point to the differences between social robots and non-social robots, and examine the relevant attributes that social robots are given, such as natural or interactive, as I survey the research of Breazeal (2002, 2003), Lin et al. (2011, 2012), Royakkers & van Est (2016), and others.

The difference between robots and *social robots* is mostly based on the fact that social robots are designed to look and behave like human entities, and to be used in human environments that require interactivity and communicability. While robots are industrial machines, HR has the strong driver to position social robots as *companions, partners,* or *assistants*. As I see it, this backdrop dominates Robot Ethics and created the concerns behind the disconnections of Chapters Two and Three. I want to explore how and why the humanoid body of the social robot became the major reference point for its agency model, how this echoes into Robot Ethics, and why computational abilities of robots, such as tracking, are essential for the robot agency, but are still ethically neglected.

The term *robot* originates from the Czech (Slavic) word *robota*, meaning 'labour doing compulsory manual works without receiving any remuneration' or 'to make things manually' (Xie, 2003: Intro). The Oxford dictionary defines *robot* as 'a machine resembling a human being and able to replicate certain human movements and functions automatically' (Intro). For Xie, the robot is increasingly living up to its name. The times in which the robot was 'merely mechanism attached to controls' (Intro) are over, due to the contemporary abilities of the robot to be manipulative, perceptive, communicative, and cognitive. Lin et al. (2011, 2012) qualify the robot as a 'thinking' machine and more than a complex of computer software. The ability to make 'its own decisions to act upon the environment' is therefore a critical aspect that separates the robot from, for example, a toaster or a coffee maker (Lin et al, 2012: 18). For them, the robot requires the ability to *sense, think, and act*. Lin et al. write:

> 'Thus a robot must have sensors, processing ability that emulates some aspects of cognition, and actuators. Sensors are needed to obtain information from the environment. Reactive behaviors (like the stretch reflex in humans) do not require any deep cognitive ability, but on-board intelligence is necessary if the robot is to perform significant tasks autonomously, and actuation is needed to enable the robot to exert forces upon the environment' (943).

They emphasise the position of the robot as an interactive, autonomous, and responsive technology. Royakkers & van Est (2016) confirm this view by quoting the International Federation of Robotics (IFR), and add that robots are, in fact, a conglomerate of machinic networks that own a certain degree of autonomy (and input sensors) while also being a social practice. They describe robots as:

> '[I]ntelligent (usually networked) machines that perform physical actions with a certain degree of autonomy within a complex and, to a greater or lesser extent, unstructured environment and a dynamic social practice. This implies, among other things, that the interaction between environment and machine, and man and machine, plays an increasingly important role. To make this interaction possible, the robot employs sensors with which it can perceive the environment and human beings' (10).

Robotics as a research field stretches back to early cybernetics and associates Robotics with Computer Studies rather than with Communication and Information Studies.[76] Contemporary research on social robots often derives from, or builds upon, concepts of human-like agenthood and autonomy, inspired by 20th century science-fiction films such as Lang's *Metropolis* (1927), or books such as Asimov's *I, Robot* (1950) or Philip K.

---

[76] These disciplines later split into further sub-divisions, such as Media Studies, Humanoid Robotics, Computer Studies/Cybernetics, AI, etc. All of them have early cybernetics as a common background (not the only one) from which these areas grew into different ways. Research foci divert by having different views on the understandings technology, human agency, or intelligence.

Dick's *Do Androids Dream of Electric Sheep* (1968).[77] Such influences are apparent in various publications in HR including Lin et al. (2012), Breazeal (2003)[78], and Wallach & Allen (2009). I noticed that literature in HR often exhibits a huge and contingent fascination for, and fear of, robots becoming truly autonomous agents, as Dumouchel & Damiano (2017) also point out, because of the science fiction background upon which many ambitions are built.

Before being included into the Human-Robot-Interaction (HRI), the figure of the *social robot* was initially mentioned in research on multi-robot or distributed robotic systems to negotiate an interaction between robots. The intention of this area was to study and simulate the collective behaviours of insects and other social animals (Breazeal, 2003; Brooks, 2002; Fong et al., 2002). The HRI research explores 'the sociality between robots and humans' (Tseng, 2016: 188) and refers to the social robot as a 'socially interactive' (Breazeal 2002, 2004) or 'Emotional Cognitive Agent' (ECA) (Ford, 2014: 27, 30). In this thesis, I exclusively refer to humanoid social robots, and not to those without a humanoid shape. However, to complicate things slightly, not every humanoid body can be understood as being a *social* device, according to HR research or automata

---

[77] For more on the history of robotics as shaped and mediated through Science Fiction films, see Telotte (1995).

[78] Breazeal takes the influence of science fiction to the point that she begins her book *Designing Sociable Robots* (2003) with this quote: 'What is a sociable robot? It is a difficult concept to define, but science fiction offers many examples. There are the mechanical droids R2-D2 and C-3PO from the movie *Star Wars* and the android Lt. Commander Data from the television series *Star Trek: The Next Generation*. Many wonderful examples exist in the short stories of Isaac Asimov and Brian Aldiss, such as the robot Robbie' (1).

history. For instance, early 18th century automata[79] would not qualify as a social robot, according to HR standards.

Despite this historical background overlapping in anthropomorphic tendencies, I was surprised to notice that early automata history is rarely referenced or known to HR research. I assume that the reason for this neglect relates to HR, in its roots, being fixated in Cybernetics and Computer Science. Therefore, it is not concerned at all with its clear contingency to other humanoid design discourses, which reach back to origins in Ancient Greece (in the context of European history) and then blossomed again throughout the 18th century with the design of 'life-like' automata (Reilly, 2011).

Automata were, in themselves, discourses on animation, agency, and autonomy of their time, and were understood as much more than an empty shell, puppet, or doll in a human-like body. According to Reilly (2011), automata are the 'precursors to our contemporary digital culture and the ancestors of the robot, the cyborg, and the avatar, demonstrating that our spectacular culture of machine-based entertainments has many historical precedents' (1). While contemporary humanoids share their appearance with their predecessors - for instance, the chess-playing *Turk*[80] - automata had a 'symbolic

---

[79] 'The word automaton comes from the Greek *automatos*, meaning 'acting of itself', referring to automated moving figures of animals or human beings. While automata look like dolls or toys, it is their animation that signifies life. This life-like movement means that automata are often perceived as if they're alive. As a result, automata are central to debates about mimesis or the representation of reality in the historical period in which they exist' (Reilly, 2011: 1).

[80] Reilly points to the multiple political and symbolic discourse that manifested in this automaton. 'While the Turk appears as a deceptively simple mechanical trifle constructed for the pleasures of the aristocracy, it is actually a theatrical object upon which the historical and discursive practices of Orientalism are staged. The automaton Turk was a *bagatelle* or playful illusion composed of working clockwork machinery: the left hand that held his pipe, the right hand that moved the chess pieces, and the noisy

and political value' (1), not a role as a companion or worker. Early automata were supposed to attract human engagement through a mix of fascination with the similar humanoid features and also by exhibiting a *lifelike* behaviour. I find Reilly's concept of the 'onto-epistemic mimesis' (7) interesting to point out. It outlines one approach through which to understand the role of perception and visual human alikeness, to comprehend how automata were understood, but this thesis argues that it can be applied to social robots as well, due to them also being humanoids. Reilly argues that the mimetic mirroring of humans in the automata's appearance…

> '[…]or representation directly shapes ideas about reality through ways of being (ontology), or ways of knowing (epistemology). In this form of mimesis, the immediacy of a way of knowing (epistemology) – or information experienced through spectacle – seemingly changes one's way of being (ontology)' (2011: 7).

If I would apply her concept to social robots, it immediately becomes problematic, since the apparent similarity leads to a neglect of the technological capacities of robots. Researchers such as Coeckelbergh (2010) critiqued this approach previously through the 'psychopathic' robot, and Johnson & Noorman (2014) urged to be mindful on what metaphoric level the similarity between technology and human agency focusses on. The reason why Reilly's concept (2011) is interesting to mention, as I see it, is that it would allow situating the use of humanoid design in a contingency to automata history, which is not a view that HR supports, but one possible if automata and social robots are

---

clockworks whirring inside his spine all provided concealment, keeping audiences from realizing that the ghost in the machine was no ghost at all' (2011: 4).

discussed in Performance Studies. What would be central in such a perspective, I think, would be the mimetic experience and anthropomorphic design that plays a huge part in the design of social robots, but ends up fully disconnected from automata history.

Despite this historical link, the contemporary humanoid social robot is the direct continuation of an automaton, so Reilly (2011) believes. She draws a clear line between robots in a wider sense: being 'workers' (6) compared to the automaton as an 'entertainer' (6).[81] Her research does not discuss the ambitions in HR to position social robots as companions, which is pushing the worker one into the background. In their new role, robots are assigned a new purpose; one of 'human companions' (Menezes et al., 2007: 367; Leite et al., 2012: 250). The new perspectives on robots as social partners have emerged because these 'can autonomously interact with humans in a socially meaningful way' (Lee et al., 2005: 538). According to Broadbent, 'an autonomous robot is a machine that can operate and perform tasks by itself without a continuous human guidance' (628).

According to research in HR, being *social* comes easy for the human, but it must be artificially programmed into the robot.[82] HR follows a constructivist approach on

---

[81] The shift from automata to robot is much more complex than can be explained here, as is the fear that comes with the focus on robots. Reilly writes: 'The automaton plays the man of the court, the socialite, it takes part in the social and theatrical drama of pre-Revolutionary France. As for the robot, as its name implies, it works; end of the theatre, beginning of human mechanics' (2011: 166).

[82] Since I support a media theoretical angle on technology and media, I would argue that the attribute *social*, which is added to the 'social robot' as a quality, is, in fact, a useless emphasis. In my view, every technology is inherently *social* (Pias, 2011; Zielinski, 2013), but this position is not common in HR, nor in POT. It is therefore important to trace what this emphasis implies for their fields, independent from my disagreeing.

making the robot a social agent, corresponding to an instrumental view on technology as neutral in the first instance and as opposed to the social human. In most robotic discourses, the *human,* as a vague concept, provides a social model from which to copy.[83] There is a tendency in HR papers to use the human agenthood concepts as undefined *social backdrops* upon which to build the social robot agency upon. This is concerning but evident, even if there are also researchers like Nehaniv et al. (2005), who critically point out that a robot is not a human, nor a computer, and researchers must therefore be careful in applying the same principles. They write:

> 'Due to the situated embodied nature of such interactions and the non-human nature of robots, it is not possible to directly carry over methods from human-computer-interaction (HCI) or rely entirely on insights from the psychology of human-human interaction' (371).

The *social* attribute attached to robots by roboticists can mean different things; from the humanoid appearance, to interactivity, and to the use in a human environment (HR does not always specify these environments, but Lin et al. (2012) and Royakkers & van Est (2016) point to professional elderly care as an important context). Hence, the social robot is distinguished, for instance, from an assembly line or war robot (Royakkers & van Est, 2016) because of a different use or appearance. Consequently, this leads to a different ethical debate around their use. I will argue in Chapter Seven that this is a

---

[83] 'Social interaction with others is a capability of humans that comes effortlessly. From birth until the end of their life, humans are constantly exposed to and engaged in social behaviour and interactions with their parents, peers and offspring. The ability to interact with groups of other human beings in a seamless and coherent manner is arguably deeply intertwined with our general development, and as a result, it has been suggested that social interaction and collaboration has also helped shaped how the human intelligence has evolved and developed over the centuries. This is known as the Social Intelligence Hypothesis' (Read, 2014: 1,2).

problem, since what a robot is used for does not relate to what computational autonomy these devices have or what ethical consequences their use could lead to. For a roboticist to decide that war robots are ethically more problematic is also correct, but only because the context of war is as such; it is not because the robot is more *immorally* acting than a social robot in elderly care.

Royakkers & van Est (2016) continue by pointing out that the physical appearance of the robot depends largely on its function (10). For them, three important features are fused into a social robot, which are:

> '[T]he physical appearance, physical handling, and observation capabilities of the robot. These characteristics are often interrelated. A certain appearance, for example, two legs, makes a particular action, such as walking, possible and other actions impossible or very difficult, such as flying' (10,11).

For the robot to act in a human environment is, for them, the minimum requirement and a straightforward argument to understand the robot as social through its placement.[84] For instance, Kanda & Ishiguro (2013) discuss the shift from industry robots into robots within 'daily environments' (2), which requires them to be interactive as another necessity. Whereby, Lin et al. (2012) expand this angle by amplifying the specific placement into private homes. They propose that 'robots will become increasingly "co-inhabitants" of humans' (Lin et al., 2012: 26), assisting in private homes, in cleaning,

---

[84] The situational view is excluded from this debate (Tseng et al, 2016), which also reflects on the specificity within the context of application (i.e. specifics in interactions with groups).

housekeeping, child care, secretarial duties, and so on. Still, these co-habitants are not like vacuum cleaners, since these do interact with people in a significant way.

For Scheutz (2012), social robots are about to become as important in society as computers already are in private homes. However, this will take time, since social robots are still a 'recent invention' (206). These will, nonetheless, become part of people's daily routines, according to Scheutz. For him, social robots are very much related to computer and industrial robots, since social robots 'contain computers (for their behaviour control) and share with industrial robots the properties of being robots (in the sense of being machines with motion or manipulation capabilities, or both)' (206). Like industrial robots, social robots have the capability to initiate motion, of others or themselves, and thus to exhibit behaviour (compared to stationary objects like computers). By comparison to industrial robots, which are mostly used and positioned in factories, 'social robots are directly targeted at consumers for service purposes' (206) and for entertainment, but the bonding will be different, since social robots will be able to imitate and interact with their 'owners' (206). This brings up my point in not having to distinguish between the social robot as companion and as tracking device, but due to whatever reason, it is not aligned in this way within Robot Ethics.

At no point do I suggest that there is a practical or possible separation in viewing the robot as either companion or computer. It is both, at any time. The concerns I have are on why the ethical robot camps do not make a clear cut on what to build the ethical

agency upon. However, this cut is not that easy to make, as Chapter Four showed and Chapter Seven will too, since it stems from different foci on agency and morality.

One trend is additionally concerning, which I want to stress, since it enhances the humanoid design towards a total resemblance with a human body, yet these new models of robots, called *geminoids*, have almost no decisive autonomy in their interaction. This implies that a human-like appearance (of the robot) is hugely favoured instead of its technological autonomy. In practice, this means that these robots are always fully remotely controlled, instead of making use of their input channels (cameras, for instance). Good examples for this development are the popular geminoids, *Erika* or *Sophia*, which are more so Public Relations (PR) stunts than functional robots. Nonetheless, the PR element cannot be underestimated in this discussion. Problematically, I would align this to what Kahn et al. (2012) were presumably exploring in their study on moral appearance in Chapter Five; one I dismissed as ontologically accurate.

Within the next section, I will explore how robot appearance, interaction, and perceived interactivity come together in social robots, and how this complicates their position as either perceived human-like agents or as computational interactive systems, which I claim are two camps of research.

On Perceived Interactivity and Anthropomorphism

Summary: In this section, I emphasise the importance of human perception and anthropomorphism in the context of HR, as I continue to show that this dynamic in HR consequently influences Robot Ethics. I provide an overview of the literature around social robots and how anthropomorphism supports their position as companions and friends. I make use of Duffy (2003), Breazeal (2003), and Lemaignan et al. (2014) to support the exploration of anthropomorphism as an important dynamic for the social interaction between robot and humans; one that is 'managing expectations' (Duffy, 2003: 178) of the human perceiver to then allow for social robots to be introduced into more environments, such as elderly care.

Making or designing technology to adapt to human-centred behaviour or interactivity might not be a new tendency, but HR positions itself within a paradigm change in the sciences. Nitsch & Popp (2014) write:

> 'With advances in psychology, neuroscience, computer science, and engineering, machine capacities continuously increase and expand. In addition, a paradigm shift is taking place from machine-centered to increasingly human-centered approaches to technological development. The field of robotics is currently particularly affected by this paradigm change' (621).

It is no surprise, in this sense, that HR aims towards making *nicer* robots or better integrating them into our daily lives. This involves the interaction with robots perceived as positive, believable (Pelachaud & Poggi, 2002: 182), meaningful, or natural (Menezes et al., 2007; Breazeal, 2002, 2003). However, this interaction (or HRI, as mentioned previously) is not always grounded on more than perceptive qualities and anthropomorphic clues. Hence, the side in HR that aims to improve the HRI is heavily invested in encouraging a perceived agenthood or companion position. One step would be the humanoid design of the robot, which encourages the human agent to perceive the social robot as a human-like (not as actually human, but neither as machine only) agent by evoking a meaningful, believable, and natural interaction between human and robot.

Anthropomorphism (not to confuse with anthropo*centrism; s*ee key terms) derives from the Greek *anthopos* and *morphe*, meaning 'the tendency to attribute human characteristics to inanimate objects, animals, or others with a view to helping us rationalise their actions' (Duffy, 2003: 180). As Duffy argues, anthropomorphism favours the observer's perspective in the interaction, thus the anthropomorphic view on

interaction. It can be understood as an 'induction' or encouragement, but not as a 'solution' towards a successful HRI (181).

The apparent human tendency to anthropomorphise human-like gestures, and therefore to treat human-like robots like humans (Breazeal, 2003: 169; 2004: xii; Broadbent, 2017), is central to debates about social robots. Duffy (2003) accentuates the importance of anthropomorphism in robotic systems to allow interaction between the robot and the human to unfold. 'A robot's capacity to be able to engage in meaningful social interaction with people inherently requires the employment of a degree of anthropomorphic, or human-like, qualities whether in form of behaviour or both' (Duffy, 2013: 178).

According to Lemaignan et al. (2014), the reason why this induction works that well is because 'people reason about an unknown stimulus based on a better-known representation of a related stimulus' and this reasoning about a non-human agent is 'based on representation of the self or other humans' (3). Hence, anthropomorphism can be understood as an interactive process between the humanoid sign (body, features, or humanoid limbs) and the human imagination to *mistake* such as representative for the recognised qualities. They refer to Lee et al. (2005) by pointing out two reasons why humans are tempted to anthropomorphise. One theory is that humans have an effect-driven, biological tendency to respond to life-like behaviour or social cues they identify in an artefact that appears human-like. This means that aliveness and humanness are conflated purposefully, because they are related in the human organism. Lemaignan et

al. (2014) add a second perspective to this dynamic from a cognitive point of view by stating:

> [A]nthropomorphism is described through people's specific mental model they construct about how an artefact works the way it does. We then anthropomorphize because it allows us to explain things we do not understand in terms that we do understand, and what we understand best is ourselves as human beings. This is consistent with the familiarity thesis. (…) people tend to thoughtfully develop a mental model of agents in their environment and make inferences about it based on what is familiar to them' (3).

There is a fine line between anthropomorphism and uncanniness, as the roboticist Mori (1970/2017) already argued 50 years ago. What Mori illustrated through the 'uncanny valley' graph is a likely negative human response to a humanoid and socially appearing robot under certain conditions. His research has influenced, and still influences, the design of humanoids (Royakkers & van Est, 2016: 10-16), even if its relevance is questioned by contemporary theories that take familiarity into account (Lemaignan et al., 2014). According to Royakkers & van Est (2016), Mori discovered that the…

> '[…]more a robot looks like a person or an animal, the more positive and empathetic feelings it will evoke in people. If robots resemble people very strongly, but their behavior is not human enough, then Mori predicts a strong sense of unease. In this case, the appearance is humanlike, but there is very little familiarity. This is what Mori calls the uncanny valley' (11).

The potential uncanniness – a graphical valley in Mori's statistic curve – leads to a 'sudden shift in our affinity' (Lin et al., 2012: 26) that immediately stops the will to interact or trust the robot. For Lin et al. (2012), the humanoid alikeness does not need to be 'completely human-like in order to be trusted by people' (26) or to form an

emotional bond. They state that a positive response from humans can be evoked by dolls and puppets, even though they do not look exactly like humans. Mori's work still explains, in a simple way, how alikeness and aliveness must correlate but not be identical, so that the robot is not seen as uncanny by the human (Ravetto-Biagioli, 2016; Duffy, 2003; Royakkers & Van Est, 2016).[85] The reluctance to respond positively to humanoid robots stems, according to Mori, from a misbalance between alikeness and aliveness.

Duffy (2003) elaborates on this by indicating that this requires finding the *as-close-as-possible proximity* in humanoid design. Hence, anthropomorphism does not require that humanoid features in robots are identical to those of human bodies or faces in order for human qualities to be projected into them. Furthermore, it seems that the very humanoid features were not that essential in the early research on this process. What this early research on anthropomorphism by Reeves & Nass (1996) shows is that when humans interact with technology, they have a tendency to anthropomorphise even computers and other technology that exhibit human-like behaviour, such as speech or gestures (more the inference of empathy in Paiva et al., 2017). However, this projection process, as they describe it, evolves from the very interaction with a technology that is responsive, and not because the computer has humanoid features by design. They create a model called CSA (short for 'computers as social actors'), a research paradigm that

---

[85] The paper on 'The digital uncanny and the ghost effects' by Ravetto-Biagioli (2016) offers an interesting exploration of uncanniness as a shifting concept, from an aesthetic experience facing a human-like machine, to the technological capacity of digital interfaces to position the human simultaneously as an agent, data, and object. I will pick this up in the section on tracking in Chapter Six to contextualise it.

'suggests that human interaction with a computer is fundamentally social and that humans apply wide sets of social characteristics to a computer when the computer manifests human-like characteristics such as language, social roles, gender, ethnicity, and personality' (Lee et al., 2005: 540).

Reeves & Nass (1996) observe that when 'it comes to being social, people are built to make the conservative error: When in doubt, treat it as human' (22). It is interesting that Reeves & Nass never specifically argued that the anthropomorphised device must necessarily exhibit a humanoid shape, but it seems such an essential factor for HR. Reeves & Nass' wider research on the 'media equation' theory[86] only ever pointed to the projection process as being essential, not the human alikeness also being such. Somehow, HR research accentuates the importance of humanoid features, which probably are meant to increase the anthropomorphic tendencies. Ultimately, the line between how human *alikeness* a shape can have before being perceived as uncanny, or how much it must be visually alike to guarantee an anthropomorphic projection, is still an undecided factor.

For Scheutz (2012), the reason why humans anthropomorphise is grounded in human biology and emotion. For him, an anthropomorphic design also links to being able to

---

[86] Gunkel (2016) refers to Reeves & Nass' work (1996) and makes a similar point, but without mentioning anthropomorphism. For him, what the CSA model brought out was that human subjects will 'irrespective of the actual intelligence possessed (or not) by the machine, tend to respond to technology as another socially aware and interactive subject. In other words, even when experienced users know quite well that they are engaged with a machine, they make the "conservative error" (…)' (242) to anthropomorphise.

create an emotional bond with an artefact, as with a humanoid robot, despite knowing that a robot cannot feel the same emotionality as humans might have for it or even see in its actions. He writes:

> 'Social robots are clearly able to push our "Darwinian buttons," those mechanisms that evolution produced in our social brains to cope with the dynamics and complexities of social groups, mechanisms that automatically trigger inferences about other agents' mental states, beliefs, desires, and intentions' (216).

Having acknowledged that anthropomorphic responses are important, it must be outlined how these are established. For him, anthropomorphism relies heavily on clues, which are ideally leading to a natural, meaningful, or moral perception of the robot, even if it is only based on a projective process; something upon which Scheutz is not fully willing to build his model agency.

Looking at the human expectations or projections of robots and Scheutz's 'perceived autonomy' offers a good way to illustrate my concerns with an anthropomorphised robotic agency in the HRI. What a perceived autonomy does is never allow for the human interactor to know how interactive or responsive a robot is. The robot appears to be spontaneously and autonomously interactive, even if it does not respond to any input or is aware of its environment, but this *interaction* originates either from a developer operating it remotely or because it is fulfilling a pre-determined program. This reduces the agenthood of the robot drastically, but it is not seen as a significant problem that needs clarification.

In this case, the robot agent position relies on what Scheutz calls critically the 'false pretence' (2012: 215) agenthood, whereby he means that a human who engages with a robot is made to believe that the robot is a present and social agent, and while it might be a present functional technology, it is far from necessarily being interactive or accurate. What this creates, exclusively, is a threefold doubt in: the robot acting on its own through its response-driven abilities (to some degree, autonomous), or what a robot can act out as it is remote controlled (to a smaller degree, operationally autonomous), and to what degree this means the robot is an actual companion or agent only because it can move its limbs.

Nonetheless, attributes such as natural, believable, and meaningful are used metaphors to increase the trust and engagement coming from human agents. However, on what grounds should this trust be built? I examine next why the rhetoric around these concepts is highly problematic and does not clarify what autonomy or interactivity means. *Naturality*, for instance, is important since it refers to familiarity, and this supports the dynamics of anthropomorphism (Lemaignan et al. 2014). Nonetheless, it remains unclear to me what research means by *natural* behaviour for a humanoid robot - a man-made, bulky machine. Mostly, what natural behaviour and expression in HRI means is that the design and behaviour of a robot must give the impression *as if* it emerges naturally within the HRI, which I think is almost an oxymoron, considering that the robot does not exhibit anything *natural,* per se. I suppose that naturality has a difficult role to play, because when it comes to Mori's and Duffy's work, the robot only

behaves/looks natural *enough* while as natural *as possible*, but must keep this balance by not becoming too close in its humanness.

The idea of natural behaviour in the robot is not a new concept, as Menezes et al. (2007) argue. In 1989, it was Engelberger (1989) who introduced the idea of robots serving humans in everyday environments. Since then, a considerable number of mature robotic systems have been implemented, which claim to be servants or personal assistants (see survey in Fong et al., 2002). The argument behind this is often that the behaviour/design of the robot must appear as natural as possible, so as to allow the human to trust the social robot as an agent (Lin et al., 2005). Clearly, after what I have being pointing out, the idea of trusting robots should not be grounded on the illusion of them judging or acting consciously.

For Kirby et al. (2010), the question of natural and easy-to-operate are bound together. To be able to understand robots means to be able to use them easily, especially in health-care environments in which non-roboticists must be able to interact with them without much training. To do this, robots are ideally responsive and display emotions such as moods, just like humans do (322), and do not require much operational knowledge.

Further, *natural* also aligns to *meaningful* behaviour, so the robot is able to manage people's expectations (Duffy, 2003: 178) while being believable – from being friendly to authoritative in conduct (Pelachaud & Poggi, 2002: 182). While there is lots of agreement on these factors being important in the HRI, De Greeff & Belpaeme (2015)

are rare exceptions in this research field. They point out that the HRI should not only be natural, 'but preferably also desirable' (26) by the human agent. What meaningful gestures are, for instance, appears to be a question of projection and expectation instead of ability to interpret (Flusser 2014), since to establish meaning in the robotic system is not only difficult, but requires clear signification. I stress that the *making of* meaning between two forms of judgement systems, human and robot, cannot be achieved simply by following an aesthetic similarity alone.

For Menezes et al. (2007), two more aspects define what *natural* interaction is supposed to be like: physical presence and predictability. He states:

> 'The first is to facilitate tasks, which involve direct physical cooperation between humans and robots. Hence, physical presence is important. (…) The second issue is that robot independent movements must appear familiar and predictable to humans. Furthermore, in order to be more effective towards a seeming interaction, a similar appearance to humans is an important requirement' (367).

According to Royakkers & van Est (2016), naturality also relates to performing the same actions as humans would. For instance, a flying robot might be strange to relate to (11), just as a robot in which behaviour and appearance do not synchronise (Mori, 1970/2017) would also be considered uncanny. This is a reason why, for instance, assembly line robots, 'Roombas' (vacuum cleaning robots), or utility robots are neither social nor humanoid, even if they fit Kanda & Ishiguro's interactive context (Lee et al, 2005: 539).

What the quote (above) by Menezes et al. (2007) quickly brushes over is the mix-up between what a human and a robot each need to perform or live up to when they

interact, since both sides have completely different starting points on what a social interaction is (the robot has none prior to being programmed). Natural behaviour also requires the agent position to be *believable,* as Pelachaud and Poggi (2002) explain in their research. They unpack this angle through what they have labelled as the 'BIEA' (Believable Interactive Embodied Agent).

Believable, for them, means that the BIEA must be able to 'manage' (181) certain aspects, such as owning information on their area of application, on the user's mental state and emotions, and on the user's intentions and beliefs, but also be capable of deciding when to provide or withhold information, since 'humans sometimes do not display their emotions' (181). The BIEA must also be able to communicate, providing the agent wants them to, and is able to express gazes, gestures, and body movements. Further, Pelachaud & Poggi emphasise the importance of the context. Their concept investigates specific personality traits of the robot. They ask critically in their research:

> 'Do we want a friendly agent or an authoritative agent? Using too simple models of communicative behaviour might produce a poor agent with which the user will rapidly feel bored. Again, some applications require caricature and over-expressive behaviours, while others need very accurate and realistic behaviours' (182).

As I already mentioned, to decide on what extent of human alikeness in design makes an 'optimal anthropomorphism' (Duffy, 2003: 182) is difficult to decide, considering that this interaction, grounded on anthropomorphisms, is a dynamic process according to Lemaignan et al. (2014: 2). Nonetheless, the relationship between anthropomorphism

and interaction is more complex and not equally positioned in terms of qualities or capacities.

There are also important voices in HR who are critical of the anthropomorphic induction as a social standard. Ford (2014), for instance, claims that humans place as much trust in automated systems as they place in humans (34) by referring to the 'automation bias' (Mosier et al., 1998). This briefly advocates that humans trust automated systems just as much as other humans, even when they are not humanoids, but it also states that when these fail, the trust cannot be easily restored again.

Breazeal (in Lee et al., 2005) reflects on how a perception-based interaction model leads to a lower expectation of sociability in the robot. Breazeal, one of the major figures in Social Robotics, is, in fact, very critical of anthropomorphism as a legitimate ontological argument in social robot design.[87] For her, the focus on such a visible design level is insufficient to tackle the question on sociability[88] and agency of robots as social agents. Breazeal considers that anthropomorphism remains the lowest form of sociability in a social robot, since it is only appearance-driven and superficial. Instead, for her, *social interactivity* must be the highest form of sociability, since this reflects the robot's abilities to be *truly* interactive, overcoming the divide of appearance/essence.

---

[87] It is hence not surprising that Breazeal's research has moved away from humanoid social robots. She is considered a pioneer in Social Robotics with developing one of the first social robots, *KISMET*. Her new creation, *Jibo,* resembles a human body only minimally. It is, rather, an advanced and autonomous home assistance system she still calls a social robot nonetheless. More on *Jibo* at https://www.jibo.com/technology/ (Accessed: 20.04.2018), and in Guizzo (2015).

[88] For Breazeal, as for most roboticists, the *social* aspect is added to the robot as an instrument; sociability is not inherent to technology.

Breazeal pushes for the capacities of social robots to become much more interactive. She justifies this with:

> 'As the most primitive social robots, socially evocative robots utilize the human tendency to anthropomorphize objects and rely heavily on users' affective responses, (…). If robots can recognize and manifest natural human-interaction modalities such as speech and gestures, they become social interface robots. (…) receptive robots have at least a modest level of social cognition, which enables them—through a simple mechanism such as imitation— to learn things (or modify their internal representation of the world) from social interaction. Nevertheless, they are not as fully socially functional as human beings. With sophisticated models of social cognition, some robots can even proactively seek social interaction to satisfy their internal states replicating human goals and desires. (…) these are sociable robots' (Breazeal in Lee et. al 2005: 538).

Damiano & Dumouchel (2018) agree that anthropomorphism and interaction are related and must support each other. Relying solely on realist humanoid design is not enough as they point out:

> 'The basic hypothesis is that strong realism *in either* of these two factors allows a robot to reach the "social threshold" where humans experience its presence as that of another social agent and are disposed to socially interact with the machine. This implies that a highly anthropomorphic robot can produce that social effect even when behavioral realism is low, and, vice versa, that behavioral realism will lead to anthropomorphic projection even in the absence of a human-like appearance. Things, however, are not quite that simple, in particular, the relation between the two factors appears to be asymmetrical' (2).

As Breazeal (2003), Lee et al. (2005), and Read (2014) also argue in their work, the interactive ability of the robot to respond to the human cannot be *perceived*; it must be built on the robot's ability to socially interact. Therefore, the introduction of anthropomorphic projection might be important, but, according to Menezes et al.

(2007), Duffy (2003), Royakkers & van Est (2016), and Breazeal (2002, 2003), it is not the only aspect necessary for a social interaction. From an HR perspective, it remains the visible, design-driven explanation of the social interaction.

After I surveyed and reflected upon why anthropomorphism is an important aspect for the HRI – but, as such, it remains problematic for a perception-based concept of agenthood – I also discussed attributes given to the social robot such as naturality, believability, and uncanniness. What I focussed on with the anthropomorphic agenthood debate was that social robots are understood through their visible *humanness* in HR, as the design aims to balance and compromise between aliveness and alikeness in looks and behaviour. As I already mentioned, reaching an 'optimal anthropomorphism' (Duffy, 2003: 182) in this dynamic process proves difficult. Still, the perceptive alikeness often dominates and drives the constitution of robot agency, as I conclude critically. Therefore, the perceptive and projected qualities do not only complicate the positioning of any agenthood in social robots, as I illustrated through Coeckelbergh's work and Chapter Two, but I laid out that this view is manifested in HR research without much reflection on what it might mean for ethical accountability questions.

Next, I review the complementing side to the anthropomorphised companion position, as I move on to explaining the computational capacities of robots ethically, which links to my concerns in Chapter Three on data and tracking. I shift to explaining the role of the computational interaction and architecture of the HRI; a view that, I think, is underdeveloped in Robot Ethics, but that is highly consequential for the ethical

dimension. However, before beginning this new exploration, I expand on two interlude sections. The first offers a different perspective on anthropomorphism, and the second reflects on the theoretical imprecision around HR terminology and ambitions, which I see leading to problematic conflations of ontological epistemes, projected expectations, and capacities around robot agenthood or agency.

On Anthropomorphism, Social Robots and Synthetic Ethics

Summary: At this point, I incorporate a brief interlude section on anthropomorphism seen in a new ethical light, as Damiano & Dumouchel (2018) suggested. Their philosophical perspective positions social robots as a new 'social species' (xii) instead of only comparing them to the human agent, which is a view not necessarily found in the major robot ethical discussions, nor in HR. As they embrace anthropomorphism and try to incorporate it into an affective and relational view on robots and humans, I consider this a valuable discussion, but struggle with fully incorporating it into my ethical discussion, since it grounds too much on how robots are perceived and responded to.

Damiano & Dumouchel (2018) offer a more progressive, philosophical angle on the social robot and its affective relationship to the human. They suggest seeing the social robot as a 'new social species' (xii), as a different kind of social agent, instead of as a human-like pseudo-agent or humanoid copy, which is either critiqued as a bad replica simulating human attributes or as deceiving human agents because of this. Ontologically seen, I agree that their view is more considered than the humanoid appearance/computational qualities, or real/fake dichotomies when simulating human attributed in robots. However, their work is purely theoretical and does not address issues beyond the agenthood discussion, as in practical accountability questions. Even if they suggest a much more fluid idea of agencies between human and robot, the ethical issues beyond anthropomorphism are not addressed, in my view.

They provide a challenging angle on HR and Robot Ethics with their suggestion of an 'Internal Robotics' research area (Dumouchel & Damiano, 2017: 121). Such a research area works against the duality of the traditional dichotomy dividing 'strong AI' or 'weak AI' (Duffy, 2003). What they suggest is that anthropomorphism should not be thought of as a deception, because this angle still reinforces dichotomies on real/fake agents, intelligence, or emotions, which is not useful for an in-depth discussion. They respond to Turkle's work critically by arguing that the human response to a robot can be *genuinely* emotional and the human's bond does not have to be any less genuine or emotional just because the robots are not feeling anything in return (Damiano & Dumouchel, 2018: 5).

The difference between Dumouchel & Damiano's (2017) view as philosophers who think about the social robot, compared to views from Robot Ethics or HR, is that they suggest looking at the fluidity between the two entities and not at the alikeness between human and non-human bodies. This is apparent through their interest in studying the affective relationship, rather that arguing about what any agent, be it human or robot, really *feels*. The 'affect loop' (Höök, 2009) is one important concept they specifically point out as useful in this debate, despite it coming from a robotics background. Yet, it suits the idea of Dumouchel & Damiano (2017) to think of emotions away from 'intraindividual' spaces (xi) and to see 'emotions '"interindividual", and not hidden from public view' (xi).

They suggest revisiting Höök's (2009) model of the 'affect loop'. In this model, the 'emotional' robot is described as owning the capacity to engage users in a dynamic interaction, which includes affective expressions and appropriate responses that trigger further reactions on the part of both the human and its artificial partner. The Höök (2009) model of the affect loop is supposed to make the user respond affectively to the robot, and involve it step-by-step towards bonding and feeling more and more involved with the system. This is 'a way that enhances the robot's social presence and favours human–robot social interaction' (Paiva et al., 2015: 1).

For Damiano & Dumouchel (2018), anthropomorphism is a dynamic that is symbolic for a 'social threshold' in the social robot. This view comes back to the required balance between human alikeness and human aliveness in the robot that Duffy and Mori refer to

in their research. This balance must be achieved so the robot is not perceived as uncanny (Mori, 1970; Duffy, 2003). For them, this balance is situated between the autonomy of a lifelike behaviour of the robot and its features of human alikeness. However, although uncanniness is a problem for them as well, Damiano & Dumouchel (2018) do not mind that the robot is being anthropomorphised by the human perceiver, since they place greater importance upon the difference between ascribing and inferring, which must be kept clearly differentiated. They argue that anthropomorphism must always refer to *ascribing* qualities, not to *inferring* ones. Hence, the idea of a perceived morality in a 'psychopathic' robot (Coeckelbergh, 2010), as morality inferred from the actions it performs, is as flawed in this view. I assume that Robot Ethics and HR are blurring the line between these two approaches purposefully.

Damiano & Dumouchel (2018) offer a discussion about 'Synthetic Ethics' (7) as a contingency of Internal Robotics, which aims to overcome questions on real or fake robotic qualities or emotions, because, as they argue, the 'simple equation between 'simulation' and 'imposture' is not only unable to account for fundamental ethical differences, but also tends to misrepresent them' (7). However, the question on how to overcome a perceived morality or other qualities remains open, as I see it. Even if they suggest distinguishing between inferred and ascribed qualities for the robot, how are these distinguished and can this be done at all?

These unanswered questions trouble my investigation. Their work does not address how the visually reinforced human alikeness of social robots will stop the human from

mistaking their apparent agenthood for an actual one. Since humans know that robots are not human agents rationally, there seems to be a fine line between why anthropomorphism allows to project human qualities onto non-human entities. To address issues arising with anthropomorphism and social robots, they promote a research area on Synthetic Ethics, since, as they say, anthropomorphism and robots will continue to occupy HR research. Therefore, anthropomorphism must be discussed further within the ethical discourse, because social robots will be increasingly used and made. They state:

> 'What SR needs are meta-level ethical analyses leading to guidelines that help it maximize the benefits and minimize the dangers of the construction and integration of artificial social agents in our social ecologies. That is why it is urgent to develop a different form of ethical reflection for SR [Social Robotics]. An ethics that shares SR's interactionist embodied approach, and, while recognizing the irreducible (epistemological, phenomenological, operational, etc.) differences that distinguish human–robot from human– human interactions, grants to our exchanges with social robots the status of a new, specific, certainly limited, but genuine, form of social relationships' (8).

While I do support their challenging of agencies or dichotomies in SR/HR, as much as real/fake humanness or what social bodies are, I do not identify sufficient answers to the questions raised about ethical issues concerning social robots as tracking devices and as dataveillance structures. My problem remains that, for this context, Damiano & Dumouchel (2018) encourage anthropomorphic tendencies and shift the ethical discussion onto how humans bond with robots. This is a legitimate way to argue, but not one that helps me to resolve the questions on moral agency or the accountability of robots.

The Influence of HR Rhetorics for Agency Epistemes

Summary: In the second interlude section, I reflect on my concerns with the rhetorical strategies in HR (and partially AI) research. I assume that the wider rhetorical imprecision I observed encourages an inconsistency in how agency models of robots are formed. I consider the discussion on agency, appearance, and capacities to be foundational for every theoretical exploration on social robots and equally influential for ethical research. What I find problematic is that the anthropomorphic ambitions in HR – to position social robots as agents and companions – are amplified by the rhetorical blurriness of projected qualities and actual computational qualities.

What I identified early on were multiple rhetorical inconsistencies within the research in HR, and their influence on conceptual views on agency. These often originate from a strategic imprecision when talking about social robots as perceived agents (taking the angle of visible interaction and perception), or their robotic/technological agency, as in their computational autonomy (taking the invisible computational angle not perceived by a human). As I showed already, the association of human values or attributes is used to upgrade robots from machines to agents. Next, I want to discuss the blurring between rhetorics and epistemes and why such does not help in understanding what agency or ethical view to study robots with.

What I notice is that wider epistemological foundations around concepts such as *morality*, *emotions*, *human,* or *agenthood* are rarely reflected upon specifically. However, these are conceptually implied within the theoretical framework in HR and Robot Ethics (Wallach & Allen try contextualising moral philosophical traditions in their work, as much as Floridi does). Most of these terms used in HR or Robot Ethics are borrowed from the Humanities, from moral philosophies, or from Psychology, and used as a supportive theoretical backbone to understand robots with. Unfortunately, this leads to new, unchallenged concerns. I believe this to be just one worrisome tendency in the discussions around wider robotic ethics.

The lack of clarity continues into the research on *emotional* robots. As I also spent substantial time surveying AE (artificial emotion) research, I struggle with certain simplifications in interdisciplinary models, such as emotion or intelligence. For

instance, in Picard's work in Affective Programming (2000) or in Ekman's concept (2003), FACS (Facial Action Coding System),[89] I could find a problematic appropriation of what constitutes emotions. The FACS allows for the creation of a detection mechanism to understand how 'people (...) express their inner states' (Read, 2014: 4), and is at the core of a simulation and detection mechanism that maps predefined human emotional expression to the corresponding emotional state. This then tempts researchers in HR to speak about 'emotional' robots (Parisi & Petrosino, 2010), despite this being an association and projection that comes from the robot being able to detect/track human expressions and de/codify these.

The problem I see here is that using a term like *emotion* (often used synonymously to affect in HRI) is not always unpacked fully and it is easy to lose track in regards to which kind of *emotion* concept robotics literature refers, and which models are used to code emotion - to be able to detect it. Scheutz (2012) identifies the same blurriness as problematic and writes:

> 'For example, researchers who work on emotions often say loosely that their robots have emotions, implement emotions, use emotions, and so on. This kind of suggestive language (e.g., during research presentations or even in published research papers) makes it easy for nonexpert readers to conflate the control processes in these artifacts with similarly labelled, yet substantively very different control processes in natural organisms, particularly humans' (215).

To some extent, this imprecision and the blurriness between the projection and the qualities might be the reasons why Dumouchel & Damiano (2017, 2018), in the

[89] The FASC will be unpacked in further detail in the section on *An Ethical View on Tracking*.

previous section, do suggest moving beyond this contradictory 'real/fake emotion' thinking in social robots. By doing so, this would enable researchers to avoid becoming trapped in ontologically problematic debates, which either consider robots as not having any emotions and as being deceptive companions, or, wrongly, as them being *emotional* companions (as a reminder, this is Turkle's major critique on the companionship position, but it might be too dualistic as well).

Even if emotions are considered as exclusively human traits linked to survival and mortality (Hay, 2014), there could be compromise found in arguing that robots are not emotional, neither are they expressing emotions to deceive*, but* the engagement of the human interactor in the HRI might still evoke very real emotions towards the robot without labelling the human agent as deceived or delusional. The fine line will be, in my view, decided by what can be expected of the robot, on the grounds of this bond.

On a wider scale, I assess that HR research tends not to pay much attention to contemporary debates in fields such as, Critical (Feminist) Theory, Posthumanism, or in New Materialism either, which negotiate links between the complexity of emotions and affects (Ahmed, 2004; Blackman, 2012; Massumi, 2002) and other concepts, such as gender, agential hierarchies, material discourse on technology (Braidotti, 2006; Barad, 2003; Parikka, 2015), and data (Pötzsch, 2017). Instead, HR research retreats often to

the Cartesian dichotomy of human/technology, cognitive/emotional, and inside/outside, and supports a conflation between affect/emotion (Gawne, 2012).[90]

However, many begin to share my concerns on this matter of the rhetorical blurriness or the aligned theoretical conflation (Hall, 2017),[91] although these voices come mainly from outside of HR research. Common critique is pointed at the imprecision (encouraged as much by the robotic industry) and the promotional and prolific rhetoric on the robotic capacities that are too easily associated with humanlike agenthood. Lipton (2017) writes in the *Technology Review* on the popular research robot, *Pepper,* that:

> 'Other robot makers skirt the issue of their machines' emotional intelligence. SoftBank Robotics, for instance, which sells Pepper—a "pleasant and likeable" humanoid robot built to serve as a human companion—claims that Pepper can "perceive human emotion," adding that "Pepper loves to interact with you, Pepper wants to learn more about your tastes, your habits, and quite simply who you are." But though Pepper might have the ability to recognize human emotions, and though Pepper might be capable of responding with happy smiles or expressions of sadness, no one's claiming that Pepper *actually feels* such emotions.' (Lipton, 2017).

I urge not to underestimate the deceptive element of a robot's abilities, which is drawn from various statements that are too easily made, since a non-expert user or customer

---

[90] Gawne (2012) points to the influence of Picard in HRI (Human-Robot-Interaction) and HCI (Human-Computer-Interaction) research. According to Gawne, Picard applies a 'quantifiable' view on affect considered problematic (106).

[91] The article on the geminoid *Erika* is very critical towards her qualities and capacities. It points to the issues in the robotic industry to be promotional and often to sound patronising and even sexist. Available at http://approximatelycorrect.com/2017/04/17/press-failure-guardian-meet-erica/ (Accessed 25.04.2018). I have no opportunity in this thesis to discuss, even if being critical of, the wider concerns that social robots raise in gender research. Especially, the work on geminoids promotes a problematic image and status of women in society that is worrisome.

could increasingly live, work, or be taken care of by those devices, yet does not understand or misunderstands their functionality or operations (Singh, 2015). The promotion of futuristic expectations grounded in nostalgic science fiction ideas of social robots creates ethical problems, which cannot be underestimated.

I think, at this point, some concerns return to the problem of media literacy, as discussed in Chapter Two, in which I pointed out that people, especially in elderly care, must be fully informed about the technology to which they are exposed. Hence, much of the advertised behaviours or functions of social robots that HR publicises through research reports are not clearly differentiated. *NAO* and *Pepper*, the most popular social robots used in research at this point, are not sufficiently autonomous in their interactive abilities. (This does not mean that they are not operationally functional; these devices can track movement, for instance, but might not move responsively.) Although they can undertake a responsive conversation with a human, they have trouble reacting to unexpected input. The roboticist Toni Belpaeme states their limitations very openly by writing that: 'We are still a number of years away from robots being able to interact with people on a deep level, but we're making progress all of the time.'[92] This might be his way of saying; responsive autonomy and visible interaction – and agential accountability – are far from being the robotic reality.

---

[92] More on Belpaeme's previous work at the University of Plymouth and the challenges in Robotics can be found at https://www.plymouth.ac.uk/news/connect/winter16/robot-home-brings-together-family-of-research (Accessed on 10.02.18).

Some terminology confusion in HR Robotics research stems from what I perceive as a dichotomous view on technology as being either neutral or intelligent (which follows a persistent anthropocentric approach). Gunkel (2016) responds to this by writing that the *two* sides in AI research complicate the discussion, even if there is not one consistent discussion on AI with which to begin. In this thesis, I outline an absurd distinction between apparent, *as if*, and actual morality in technological systems. This conflict is situated in AI research but impacts upon HR. In this context, there is a differentiation between 'weak AI' versus 'strong AI' (Duffy, 2003).

On the one hand, the research has not progressed as far as expected in terms of creativity or the ability of the robot to make judgments as an *intelligent* machine, which are qualities still lacking in machines (defining the deficits depends on how *intelligence* is conceptualised). Nonetheless, on the other hand, AI has created learning systems that can 'make decisions and take real-world actions with little or no human direction or oversight' (239). Gunkel quotes what Winograd (1990) wrote almost 30 years ago by writing:

> '[A]rtificial intelligence has not achieved creativity, insight, and judgment. But its shortcomings are far more mundane: we have not yet been able to construct a machine with even a modicum of common sense or one that can converse on everyday topics in ordinary language' (Winograd, 1990, quoted in Gunkel, 2016: 239).

This returns to one fact: making a difference between an intelligent system and an *apparently* intelligent system is not an unsubstantiated one, since what is *apparent* intelligence is measured up with human intelligence, not algorithmic intelligence. I

notice further that, from the human perceptive on the perceived level of technology, it is hard to distinguish between a robot being intelligent or one only *appearing* intelligent, since the human perceiver never really knows who or what controls the robot and whether it is interactive.[93] And yet, some levels are totally invisible. The problems I see here are, therefore, not only on the anthropomorphic induction being deceptive, as Turkle (2005) argued, but also in statements like Duffy's (2003) on the robot being able to 'pretend to be intelligent' and to 'cheat' (180), since both mistake inferring intent with having intent.

I would like to explain this implication in more detail. I would say that cheating requires an *immoral* intention and the ability to distinguish between morally *good* and *bad* behaviour. However, it is still not technically possible to reproduce or simulate human reasoning in a robotic system, so that these devices can perform such a sophisticated reflection process. The question instead always comes back to: Can the robot execute man-made implemented rules or not? This might imply that terms such as intelligence, morality, and other discursive concepts are reduced to a set of rules, (which might be reductionist, but at least possible) and nonetheless, these rules must be *executable* in their very reduction.[94] Hence, to correlate the questions on intelligence or morality to appearance means to correlate these to human alikeness. Aiming for this will always

---

[93] It is impossible to distinguish between these two terms, as illustrated in Bruce's et al. (2000: 4002) work on robots used in drama. Here, they first argue about an importance of AI for a believable, autonomous personality of a robot with emotional responsiveness. They then conclude that, due to the complexity of emotions, the representation of 'emotional behavior informed by biology and psychology is an important goal in agent research, [but] it remains a distant one at best' (4002).

[94] More on this debate is well documented in Ward (2015), who explains the conflict between 'weak' and 'strong' AI, relating to how intelligence evolves into morality.

raise the critique of the robot being a *worse* copy of a human agent, but ignoring its own agency.

Dumouchel & Damiano's (2017, 2018) philosophical re-figuration towards 'Internal Robotics' and 'Synthetic Ethics' offers a new angle from a techno-philosophical camp that tries to overcome the appearance/essence dichotomy in social robotics - the *as if* trap - as much as the already established research path towards 'strong AI' (Duffy, 2003) or 'new AI' (Cañamero & Lewis, 2016). As I already stressed, I am doubtful of the practical realisation of their theoretical concept, since it remains problematic to align such to the HR mindset. As pointed out several times, I strongly advocate for thinking beyond an anthropomorphic ethics exclusively.

I now move to the complementary discussion on social robots as companions; the computational interactivity within the robot's *inner shell* and beneath apparent or perception-based discussions. I will examine the computational agency and tracking interactivity of social robots in the next section and state how this capacity contributes to their agency and morality discussions essentially.

On Computational Interaction of Robots and Tracking

Summary: Within this section, I focus on explaining the computational structure and architecture of social robots, as discussed in HR and in wider Robotics and mainly supported by Rossini's work (2012) on gesture recognition. The assigning of agenthood to social robots requires a wider understanding of computational autonomy, one I continue to debate in Chapter Seven, when the moral agency of algorithms re-enters the discussion in more depth. I increasingly link data ethical concerns (Chapter Three) to the computational sociability (Chapter Six) to algorithmic morality (Chapter Seven), so as to illustrate that each of these themes operates with its own agency model, and that none of them are connected ethically yet, while all of them circle around the social robot's ability to track, collect data, and to interact, in my view.

Almost twenty years ago, Billard & Dautenhahn (1999) predicted that social robots as embodied agents in society will be accepted when they learn to recognise one another, engage in social interaction, and explicitly communicate with and learn from one another. If this were to be possible today, then social robots would deserve the status of agents, but, for now, this remains a fantasy. As I reflected on in Chapter Four, artefacts have inherent agency, but not necessarily a moral one (Johnson & Noorman, 2014), and as such agency may increase, so could their computational autonomy (Wallach & Allen, 2009).

As I moved on in this thesis, I contextualised my critical perspective on the anthropomorphic agenthood position of social robots and increasingly shift from an anthropomorphic critique towards an anthropocentric model, which will prove problematic as well. As I showed, HR is a field that addresses social robots from their perceptive dimensions, but does not ignore their computational qualities either. The discussions are fragmented, instead, into sub-research, such as HRI or gesture recognition research, which I will explore next.

I stress that building expectations for social robots to be accountable agents is even more worrisome, considering their increased use in elderly care and their practical ethical consequences that I highlighted in Chapters Two and Three. There, I was concerned with the perspective on robots' human alikeness or human sociability. However, as I pointed out, looking at robots only as humanoid bodies does not support

an understanding of their interactive abilities, which extend beyond their visible features.

Lee et al. (2005) support my critique by consequently stating that a social robot needs computational 'capabilities of conducting a wide range of social functions [e.g., speech recognition, speech generation, visual recognition, affective responses, turn-taking (interactivity), and artificial intelligence]' (540). Hence, laying out the processes involved in and required for the computational interaction in the HRI are, in my view, enormously important to understand robot agency ethically. The reason is an obvious one: Computational autonomy of robots allows for robots to navigate through a human environment and with a human subject or other objects (Menezes et al., 2007: 367).

The concept of the natural interaction, which I previously discussed, does not only need to be perceived, but it also links to a system running in the background that encourages a real-time responsiveness. According to Ziafati: 'Robots that are supposed to interact with humans have to process a great deal of information very quickly and adapt their behaviours according to the interaction' (2016).

Royakkers & van Est (2016) emphasise the technological complexity of robots as interactive systems, and encourage looking at their complexity as computational systems and not just at their perceived humanlike agenthood. They point to the multiplicity of networks and systems at work in what is understood to be an *entity* and state:

> '[T]he modern robot is not usually a self-sufficient system. In order to understand the possibilities and impossibilities of the new robotics, it is important to realize that the service robot is usually supported by a network of information technologies, as is, for example, the Internet. Thus, this implies, in particular, networked robots' (11).

In this quote, they point to two important aspects. First, robots are part of a 'dynamic social practice' in which the human and robot must learn to communicate. Second, the robots contain a *microcosm* of networks of information technologies, programmes such as Python or C++ (Read, 2014), and, often, at least one sensory (visual) detection system, such as a camera (Brèthes et al., 2004; Fong et al., 2002; Royakkers & van Est, 2016). It makes sense to diverge into a computer scientific discourse to understand how computational interaction works, in terms of how the computational architecture understands interaction between modules.

I will focus on Rossini's (2012) work in this section, so as to provide an overview on how gesture recognition modules function within the wider robotic system. Gesture recognition is a good example, in my view, through which to illustrate tracking and social interaction on a computational level. What Rossini (2012) introduces is, in fact, an illustration of *tracking* as a process, but he does this on a wider level, as computational architecture. He compares the robot with other Embodied Conversational Agents, so-called ECAs, which do not face the same difficulties in engineering as robots do. Rossini focusses on the understanding of gesture simulation as essential for the sociable engagement between human and robot, the so called HRI. Both Read (2014) and Rossini (2012) mention the use of C/C++ programming language, which operates in two main sub-systems. They mention the necessity of 'the parser' that allows for the

recognition of the users' speech, gestures, and sometimes facial expressions, and 'the planner' that allows for a response by the robotic agent (152). The parser is based on a model of the user - in this case, a human input, in the form of gestures and expressions. This requires the robot to have a social model of interaction embedded in it, before it is able to recognise or interact with a human. This is possible because of the parser being fed with an embedded 'cognitive and behavioural model' or 'some sort of emotional intelligence to allow for emotional recognition' (152).

However, programmers must first agree on what constitutes a *human*, a *gesture*, or *emotion*, to then translate norms and specific parameters into code, which is a complex computational process in itself that, ultimately, leads to the design of detection and tracking modules. What is additionally important is the synchronisation of modules with the locomotive system, and the coordinating of the robotic limbs is extremely difficult to execute. Hence, even if a tracking or detection module is functional and able to gather and decipher input, it still requires an operational connection and synchronisation to other modules. Only if this works can a corresponding, physical behaviour from the robot follow.

One major challenge is, therefore, to synchronise the input that the robot receives from its environment with the already embedded modules. The ECA model requires, according to Rossini (2012), a perfect linkage between its knowledge base, which is a set of models and pre-determined information about what is to be detected, and the information derived from sensory input. The robot's biggest challenge is, therefore, to

process ‘the information from the outer world by means of its sensors, to analyse them through the Knowledge Base module, plan a response, and execute it’ (Rossini, 2012:153). Computational architectures of robots can be distinguished as *function-based* (Nilsson, 1984) and *behaviour-based* (Brooks, 1991) architectures. According to Rossini, *function-based* architectures operate in a linear path of rules of execution, while *behaviour-based* architectures are composed of *response modules* that react to the environment without requiring a planning stage.

Next, I would like to particularly focus on how to understand tracking modules, which can be represented as one module in the computational architecture. Tracking is not an exclusively robotic or computational capacity, but a process to detect and decipher information broadly speaking. To *track* means to locate, to trace, to contextualise a position of oneself or oneself within the environment and in nature (Bray, 2014), as much as it relates to the regulation of infants’ attention (Tomkins, 2008). At the beginning of systemic tracking through mechanic tools, different civilisational steps were passed through to enhance the personal orientation and the ability to ‘read nature’ as, for instance, ancient Polynesian cultures did (Bray, 2014: 3). However, in this thesis, I am only concerned with the complexity of digitally-based tracking systems, even if tracking systems are used as much in personal computers or other devices, such as CCTV cameras.[95]

---

[95] Contemporary popular tracking applications often relate to web browsers and applications such as *cookies* (Acar et al., 2014; Raley in Gitelman, 2012). These are explored in depth in HCI/surveillance research, due to conflicting interests of companies gaining user data and the users’ privacy protection.

For the social robot to interact with a human or the environment, it needs to, at least, be in the position to detect, recognise, and to respond to human input; to recognise human shape; to process data of the embedded emotion or gesture concepts[96]; and to create meaning holistically. Hence, to be interactive or social, beyond anthropomorphic designs and human response, the social robot needs to track its environment and to process the data. To be able to track, computer vision is an essential quality, according to Royakkers & van Est (2016). For them:

> 'Robot technology progressively masters more and more complex operations. This is made possible by improved visibility (via 3D vision systems), better navigation and mobility, better voice recognition, and smarter interaction with people' (3).

In *A Survey of Socially Interactive Robots* by Fong et al. (2002), as one of the early and still relevant papers on social robots, they point out how important it is for social robots to be able to track people, gaze, and speak. They specifically use the term 'tracking' (154) as a robotic ability to locate and detect human bodies in a space in front of the robot.[97] HR and Robot Ethics limit the view on tracking to its functional and operational role in the framework of HRI, as I briefly outlined in Chapter Two, but it also allows for dataveillance to be an ethically problematic consequence, as I reiterated in Chapter

---

[96] From a media philosophical tradition, gestures and emotions are concepts which are more than anthropomorphic sign pattern. In some philosophical debates, which I cannot fully explore here, gestures are much more than visually coded and anthropomorphic actions, because the idea of gesture does not align simply and causally to a signified expression, but emerges from a bodily mediality and movement (Ruprecht, 2017; Agamben, 2000; Flusser & Roth, 2014).

[97] Tracking systems focus on capturing different human features. For instance, on the face in 'face-recognition systems' (Moubayed, 2012), the eye or gaze in eye/gaze-tracking systems (Prakash et al., 2016; Jokinen, 2009; Admoni & Scassellati, 2017; Moubayed, 2012), or on hand and eye tracking systems (Brèthes et al., 2004, 2005), while noticing tendencies in HR towards an increasing focus on the design of iris recognition systems (Jain et al., 2012).

Three.

Durantin et al. (2017) write about tracking as being the robot's social skill: 'The social skills required for a social robot include detecting, creating, and learning the meanings of social moments' (1). Equally, Breazeal et al. (2000) emphasise the importance of common perceptual abilities between human and robot, less as identical traits but more as common nominators. They state:

> 'One of the most basic is that robot and human should have at least some overlapping perceptual abilities. Otherwise, they can have little idea of what the other is sensing and responding to. Vision is one important sensory modality for human interaction, and the one we focus on in this paper. We endow our robots with visual perception that is human-like in its physical implementation' (1).

Tracking software or programs is therefore *implemented* into robots and becomes part of their 'network', as Royakkers & van Est (2016) would say, but this does not mean that they are necessarily discussed as exclusively robotic applications. To understand tracking fully and as an ethical network, I suggest discussing it through an ethical lens, after having established a basic understanding of computational architecture and modules in this section.

An Ethical View on Tracking

Summary: As Chapter Three pointed out, an ethical understanding of social robots includes taking data concerns seriously and understanding tracking as an essential component of social robots. I suggested grasping social robots amorally and without evaluations of good or bad agents or intentions and, instead, perceiving them as relational networks of intentions and capacities. In this section, I expand on an alternative ethical view on tracking, from the perspective of its non-hierarchical unfolding. I then revisit Agamben's (2000) concept on the ethical gesture to highlight that the unfolding of tracking modules (Haritaoglu et al., 2000) and the problematic but influential FACS model by Ekman (2003) are ethical networks as well. I undertake this speculative exploration to offer an alternative path to Part Two's morality-bound focus and the anthropomorphic view on robots, since these are not aligned with data-related issues. I emphasise, again, that this compromises the view on tracking, data, and robots as ethically challenging technology.

To understand the technological and ethical entanglements within tracking as a reciprocal and ethical process, the computational architecture previously unpacked presents only one way; looking at the conceptualisation of tracking modules and the concepts that fuse into it would be another.

I propose looking at the 'W^4' (809) model by Haritaoglu et al. (2000) to reflect on the unfolding of tracking modules as a synchronisation between behaviour-based or knowledge-based information. I only illustrate a few steps – from when an input signal is captured to how is it decoded, classified, traced, processed, and stored – through the W^4 model, which provides a simple explanation for understanding how multiple steps fuse into *tracking* as a summative term of this simultaneous unfolding. Haritaoglu et al. (2000) amplify the importance of visual input and its processing, and write:

> 'Visual tracking is one of the most important fields of dynamic computer vision and it provides fundamental technologies to develop real world computer vision applications: human tracking and identification, intelligent transportation, traffic flow measurement and object tracking in smart rooms' (809).

The detection process is the first step according to Haritaoglu et al. (2000). The W^4 model includes a statistical-background model with a collection of 'foreground blobs' (809) that first detects such blobs in order to then be able to put them into predetermined classes by undergoing a 'silhouette analysis' (809). These silhouettes would be defined broadly as human, group of humans, or other objects. They continue with:

> 'If a blob is classified as single-person, then a silhouette based posture analysis is applied to the blob to estimate the posture of the detected person. If a person is in the upright standing posture, then a further dynamic periodic motion analysis and symmetry analysis are applied to determine whether the person is carrying an object. (…) If a blob is classified as an object other than a person, W^4 does not do any further silhouette analysis; it simply attempts to track the object through the video' (809).

After the silhouette-based analysis of the blob is completed, a tracker calculates the correspondence between previously tracked blobs and currently detected blobs. This is how the tracking function constructs or calculates an 'appearance and motion model which enable to code and recover the trajectories of the tracked blob' (809). According to Haritaoglu et al., there are more steps that follow, but these, for now, are enough for my intention to outline the intersection of different steps merging in this process, which then fuses into what I refer to as *tracking*.

These various steps and different classifying patterns and models are man-made concepts of how a human is distinguished from a *blob* or an object. One could argue that these algorithmic *collaborations* between the human input are what constitute the computational autonomy of the robot, as discussed later with Wallach & Allen's work (2009) on computational decisions and moral agency. The synchronisation of instructions requires the architectural preconditions, which align the instructions with the processing abilities to then allow for a functioning tracking process.[98] I summarise the concept of computational autonomy only partially to highlight the *mediality* and

[98] I will not expand on the material dimension of algorithms, but I consider the algorithmic architecture as a progressive discussion on new agencies and aesthetics (Parisi, 2013).

unfolding of tracking in the tradition of Agamben (2000). This should allow exploration of the question: How does a tracking module know what to track, ethically?

As Wallach & Allen (2009) will point out in Chapter Seven, the engineer (or developer) has the responsibility of implementing values and norms, but also makes use of disciplinary models in this process. The formation of modules is not a linear or causal process, neither is it a neutral one. Even before considering the engineer's *moral* values that influence the module's rules, according to Wallach & Allen (2009), the conceptualisation of emotion or gestures must be seen as an already ethically charged process that is strongly influenced by disciplinary frameworks.

In the context of gesture recognition, Rossini does not unravel what kinds of models are used to program the modules, but he mentions the relevance of modules that process 'Emotional Intelligence' (EI or AE). An EI module is for Rossini (2012):

> '[R]esponsible for the internal state of the robot: this module can be either juxtaposed to other modules in the architecture in the creation of a *social* robot, or be included in the decision making module of a deliberative-reactive system' (154).

According to Menezes et al. (2007), the tracking module needs a concept of what a gesture or emotion is, but, as Haritaoglu et al. (2000) already mentioned, it takes certain steps from *blob* to *human* and therefore from movement detection to emotional expression. When researchers such Adascalitei & Doroftei (2012) argue for the 'emotional expressions' of robots, they often refer to Ekman's *Facial Action Coding*

System or FACS[99], which is a detection scheme used to code and decipher emotional expression in humans. The FACS is used for robots and the HRI, but also for Human-Computer-Interaction (HCI) equally, and is applied by social roboticists such as Breazeal (2002, 2003), Paiva et al. (2017), Leite et al. (2012) and Kanda et al. (2013: 304).

I debate the ethical role of the FACS in the wider framework of tracking processes next. Unpacking its conceptual constitution and intersections should highlight the importance of the epistemological context, out of which algorithmic modules are designed and concepts are embedded into robots, as much as it amplifies disciplinary goals in Robotics or Affect Sciences. This means, when I talk about gesture or emotion tracking modules in social robots, I also acknowledge that what a *gesture* or an *emotion* exists as is not a given concept, nor a stable one. In some cases, the conceptual models, which are embedded in the computational modules, draw upon earlier psychological or communicative models (as in gesture recognition) in wider communication studies research (Kendon, 2004; Knapp, 2014, Ekman, 2003).

Similarly, so does the FACS; a scheme is implemented into robots/computers to detect and decipher (around forty schematised) human emotional expressions, and these are aligned to the emotional *internal*, as Ekman would claim. This system is still incredibly

---

[99] 'Ekman is best known for his work on deception detection, the influence of which has extended beyond the academic field of psychology to the development of police and military interrogation techniques. In the 1970s, Ekman and his colleague Wallace Friesen undertook an eight-year-long study of facial expressions, creating a scheme of forty-four discrete facial "action units"—individual muscle movements combinable to form many different facial displays' (Gates, 2011: 22).

influential in the development of facial recognition software, because it allows for the standardising and causality/inference steps in detecting the so-called *basic* emotional expressions in peoples' faces.[100] As Gates (2011) argues, the FACS is still the 'gold standard' system (22) for the detection and coding of facial expressions for the HCI (Human-Computer-Interaction) and the HRI (Human-Robot-Interaction). What the FACS has conceptualised to being an *emotion* relates to the purposeful reduction of emotional expression to forty different facial expressions, which feeds back into what is trackable as emotion and what is not.

Even if Ekman has created this system from the perspective of a psychologist, the FACS was not designed to understand human emotions as I see it, but to make human emotion more readable for a visual recognition system. Hence, it aims for the signification and conceptualisation of emotions into tracking modules or units by programming systems, and must be understood under the premise of increasing the readability of emotions for technical systems, not as an attempt to understand human emotions. I want to clarify that the robot does not, in fact, *track* the actual or expressed human emotion, but that the given model manoeuvres in an outside world, corresponding to the knowledge base given to the robot that, in turn, looks out for a visual input on what emotion must look like. If I relate this dynamic to Agamben's (2000) suggestion that cinema is gestural because it remediates the visual input through its own technical sophistication, then robotic tracking does the same as cinema

[100] Breazeal also used the FACS in her pioneering social robot, *KISMET* (Breazeal, 2003).

montage; it intertwines various agencies, creating a new reciprocal bond between technological and human into an intertwined ethical structure.

Knapp et al. (2014) and Gates (2011) raise their critique on the FACS and the wider issue of standardising emotions and affective responses that I share. The FACS is a model that is culturally negligent to the differences between cultural norms and facial features that go beyond the Western (White) face. Its standardisation of emotions further excludes less expressive faces and people, by implying that a less gesticulating or less expressive person is less emotional, which is not only a dangerous assumption, but one with a biopolitical weight. Knapp et al. (2014) argue that:

> 'Although the same potential for showing a particular facial expression of emotion may exist in all humans, such as with anger, cultural upbringing influences when and how it is shown (…). Another example concerns grief. In one society, people may weep and moan at a funeral, whereas in another they may celebrate with feast and dance' (258).

This aspect illustrates to me that not only is the FACS a limited and reductionist scheme, but its reductionist modelling mostly suits the robot's processing capacities, while not allowing for any subtlety or ambiguity of human emotional expressions to be captured. Even if designed from an anthropomorphic and psychological point of view (Ekman, 2003), it represents much more than a technological understanding or pattern making of emotions as opposed to emotional expression or human emotion.

Furthermore, I stress that the cultural standardisation of emotion is a biopolitical process that aligns tracking to the biopolitical value that early cinema had, according to Agamben. The on-going standardisation and schematising of emotions and care cannot

be underestimated biopolitically. According to Gawne (2012) and myself, that aligns tracking in a wider biopolitical dispositif of emotion detection software and data gathering ambitions (and their commodification).

The further alignment to interdisciplinary scientific objectives confirms the importance of digital signification of gestures or emotions. Just how consequential and far-reaching the FACS is shows in disciplines such as Affective Programming/Computing (AP) (Picard, 2000), a field heavily influenced by the groundwork of Ekman and his FACS (Angerer & Bösel, 2015: 52) that extends the biopolitical consequences of tracking to the quantification and growth of affective technologies (Angerer & Bösel; 2015; Gawne, 2012). For many years, AP has found great resonance within the research of the HRI, but especially in aligning robotics research with the HCI. Picard, who pioneered this field (and worked with Breazeal, the Social Robotics pioneer at this MIT), explains her research aims in her book, *Affecting Computing* (2000):

> 'By coupling affective pattern recognition with wearable sensing, we have a new opportunity to teach a computer to recognize the basic affective responses of its user—for example, if the user likes or dislikes something, or is confused, or frustrated—without the user having to explicitly explain this to the computer' (250).

These models, patterns, and concepts are never neutral or universal, but become problematic when they drift into a reductionism, or align increasingly with quantification schemes and reduce the dynamic and reflexive structure of terms such as emotion or affect. This reduction can be illustrated by looking at the influence of epistemological models on affect or emotions as being intertwined and reciprocal

concepts for tracking.

In my ethical exploration, I also consider another interesting aspect. Not only are the decisions on *which* models to embed within tracking systems significant, but, furthermore, which concepts and frameworks are *not* used to define emotion or tracking, which might also be important.

For now, I did not look at the implications of tracking, but only to its formation. For instance, there are many more ethical strings that exceed the process but attach to disciplinary goals. In the wider picture of using tracking and social robots, none of the decisions made towards the design of tracking modules is neutral, causal, or non-ethical, since the various intertwined considerations influence relationships, concepts, and disciplines, resulting in far-reaching biopolitical consequences. Looking closer at the FACS highlighted to me that not only does tracking influence the coding of emotions, but that this process has allowed for emotions to become a data-based currency in the wider context of 'affect economies' (Angerer & Bösel, 2015). This is not to be underestimated if applying this realisation to elderly care and social robots.

What the research on tracking and emotion detection also displays is a theoretical dispute between disciplines. There seems to be an increasing gap forming between Affect Studies[101] – associated with the Humanities/Cultural Studies/Posthumanism and

[101] Affect Studies as aligned to the Humanities (Massumi, 2002; Gregg & Seigworth, 2010; Angerer et al., 2014) should not be mistaken for Affect Sciences as aligned to the cognitive sciences and cybernetics (Davidson et al., 2003).

Affect Sciences – associated with HR/AI/AP. The research in Affect Studies on emotion and affect, even if appraised in its own framework, is mostly disregarded in Robotics, HR and AP, which are affiliated to the theoretical concepts of emotion in the Affect Sciences. Research from Affect Studies does (mostly) not provide any quantifiable theory on emotions and affect, and seems *irrelevant* for HR or AP to be able to appropriate concepts from. However, it is the Humanities-led discussions on affect and emotion that allows for a holistic ethical discussion on tracking as aligned with Posthumanism or Agamben's work, not the one in HR or AP.

Affect Studies are overlooked in wider scientific and robotic discussions and research on artificial emotions; on the contrary, Affect Studies are very much aware of AP discussions and their relevance on shaping emotion research, as Gawne's (2012) work already showed. Gawne formulates a critique on Picard's work conflating emotion and affect as a strategical approach of the Affects Sciences, one that is not neutral, nor innocent. Such an overlap in terminology allows to conceptualise the 'one-to-one affective relationship' (106), which is Picard's intention. The problem would not be that these terms cannot be interchanged, but that this strategic 'slippage' (106), as Gawne sees it, opens the door to the appropriation…

> '[…]of affect as a quantifiable substance, which can be measured, interpreted, learnt and directed. (…) Thus the potential for opening a deeper affective engagement within the confines of this informational model is limited by this reduction of affect to a quantifiable unit. This, in fact, has reproduced some of the very problems that advocates of affective computing had identified and critiqued in the cognitivist approaches to artificial intelligence' (106).

I resume from my exploration that deciding on the concepts within tracking modules is deciding on the standardisation of emotions to specific physical expressions, and not only on robotic behaviour or gesture recognition, but also on human physicality equally. Trackable emotions are the result of various decisions and choices in the epistemological framework that fabricated their parameters. Therefore, I do not see a possibility to have a discussion on moral agency out of these realisations, since the embedding of these various agencies involved in tracking can be seen as a reciprocal and relational, hence ethical process, as Agamben would have argued.

However, the decision to evaluate these steps (even as outcomes), or the attempt to assign accountable agenthood or agencies, would not be possible, since where should one start in the chain? Can Ekman be held accountable for social robots? Or the programmer implementing the tracking module? And what if the modules are designed with a *bad* intention; are the robots then the *bad* agents?

If I revisit Agamben's analysis on cinema, then tracking systems would not simply be the linear result of Picard's or Ekman's concepts, but would be built out of the alignment and co-shaping of constant input and remediation. Tracking would be, in his sense, always gestural, just as cinema is a remediation of a representational crisis of gesture and its transformation through a new technology. I think that this analysis justifies recognising the wider spectrum in which tracking operates, beyond being a gesture or emotion recognition software, but as, in itself, a *gestural* process.

What emerges from this discussion is, in my view, that these points enable me to have

an informed and contingent understanding of the agencies involved in the complexity of processes such as tracking and, on a wider scale, of robots. This does not allow for finding moral agenthood in robots, modules, or computer systems. I see a problem however, in how expanding such a network could go too far in tracing, but without reflecting on the constructivist terminology, which then would hinder a better understanding of robots - especially if these are labelled as emotional or moral without much criticality.

I will try linking this ethical view to a better understanding of dataveillance as an ethical problem of robots. What can be drawn from an ethical discussion on tracking for the ethical consequences of dataveillance? I might not have resolved the ethical concerns, but I advocate that such an ethical discussion on tracking allows for *seeing the bigger picture* on socio-technological agencies in this process, and how this is suited for a theoretical understanding of how tracking works.

However, I am aware that this does not fix or conclude the accountability question on social robots. It should allow us to see more concerns and create an uneasiness nonetheless, and allow for second-questioning their infiltrating use in elderly care. What emerged is that, without any doubt, there is a lot at stake when considering the robot as a *bad* companion. The wider ethical consequences, to me, seem to increasingly emerge as pressing and unresolved. What I consider insufficiently stressed in these debates is that the ethical dimension in using social robots must look at the deliberate decisions of

a human agent (be it a programmer or robotics as a field) to apply tracking technology into sensitive environments.

I state that continuing a discussion on dataveillance as part of a mediatisation process of elderly care could, consequently, take two paths. One would be to perceive dataveillance as a way of distributing and accumulating the agencies that tracking supports. This view would not reflect on accountability of robots as dataveillance and informational structure, neither would it blame the human agent, or the developer, to exploit people's data, since what this perspective offers is to look at the width of and intersections of relations and the horizontal alignment of agencies. This would be what Posthumanist theorists and Agamben pursue with their ethical approach.

If I look at dataveillance as a problematic process due to the data-related magnitudes, which it raises as an informational structure, it is not an unproblematic tracking process, since it raises the question as to why there is not more reflection on the human agent who is deciding on the implementation and expansion of technology into more environments, and on the possibility to use and reuse data for other purposes than those for whom it was collected (Püschel, 2014). I would emphasise that to understand dataveillance as being ethically problematic, the human agents' (programmers, developers, roboticists, etc.) decisions to gather data, and the contexts of gathering such data, as much as the topic or context on which data is gathered, must be focused on much more.

Therefore, what the ethical discussion in Robot Ethics and POT must highlight vehemently is the importance of the human agent (developer or roboticist) as a crucial threshold that elevates data tracking to dataveillance as being *the* agent who manages the agencies involved in this process. Agamben's ethical work does not offer enough to do this, due to his focus on the entanglement of reciprocal agencies and the dissolving of boundaries, as well as the remediation of private human expression through technology not allowing to address any accountability questions, even though he labels this process as biopolitical. Neither are the Posthumanist theories discussing (practical) accountability fully, which would allow addressing questions on the consequences of dataveillance for elderly care.

I acknowledge some deficits in the exclusively ethical discussion nonetheless, as I realised that dataveillance is more complicated and comes with a negative connotation, since it implies the negative consequences already in its ethical structure. The Posthumanist ethical view might be limited in reflecting on dataveillance, since the moment in which an ethical conflict emerges, such as data infringement, the possible and practical consequences of tracking are not graspable and accountability is not a key concern. A mixed framework of moral-ethical questions, as Brey (2014) suggests in Chapter Seven, might be more suitable to tackle these.

The concern around the potential of dataveillance stems, for me, not only from its ethical unfolding, but from its negotiation of agencies and its sophisticated computational autonomy, which appears to be increasingly uncontrollable and

intransparent. Its problematic ethical consequences are not immoral decisions of the robot, or immoral intentions per se, but are, in fact, decisions leading back to human interest, human ignorance/negligence, or intentions of commercial and research independent drivers.

Dataveillance needs this extra decision that only a human agent can make; to extend tracking structures systematically into further environments. This second perspective allows us to understand dataveillance as both *ethical* and *ethically problematic* at once. The difference between this perspective and the moral discussions would be that this view does not diminish the complexity and agencies manifested in dataveillance as being still ethical; neither does it consider anything in this structure to be *good*, *bad*, neutral, or morally accountable. Considering the impossibility to demand any remorse, responsibility, or accountability from technology, practically speaking, these moments and decisions in which human intention is key must be recognised and questioned with a greater scrutiny.

Ultimately, the ethical exploration has enabled me to create a bridge between the biopolitical consequences of tracking as a process, which not only shapes the definition of emotion, but also the value from its quantification through data gathering technologies. The application of social robots has even more ethical and biopolitical consequences considering their use in elderly care. However, aligning this process to social robots seems an underexplored connection.

I claim that it is important to look at social robots beyond the reciprocal elements of tracking modules to grasp the decisions and moments required to enable dataveillance.[102] To do this, social robots cannot be limited to their perceived companion position that HR and Robot Ethics gives them, but must be discussed as a digital device that affords affective relationships from the human subjects with which it interacts. The increasing use of digital technologies in elderly care does not only support a mediatisation of care (Lundby, 2014) by reinforcing a ubiquitous technological structure, as Andrejevic (2012) mentioned, but also reconfigures the meanings on social interaction, professional care, the value of elderly people in/for society, the interest a society (roboticists, computer scientists) has of elderly people, or the interest companies have on the elderly. As it seems, these devices will additionally change the view on human caretakers (a perspective I had to, unfortunately, neglect due to the focus of this thesis) and on how they are trained, valued, and also paid, in that the use of robots will transform the care profession completely.

Beside the influences on professional care, I identify that the robotisation of labour[103], aligned with the biometric calibration of emotion, is not only relevant for AP as a research field, but is also relevant to various affect economies (Angerer & Bösel,

---

[102] To remind the reader on the requirements of dataveillance: One is the shift from top-down strategic surveillance towards ubiquitous and increasingly autonomous technology environments, then the shift from requiring an *apriori* intention on what data to gather, and third, the ability to appropriate the gathered data sets *posteriori* and to re-use them in new contexts.

[103] This refers to the increasing use of robots for labour environments in which human labour is dominating for now. Banking and postal services are already affected by this process in which the human labour is increasingly less required, due to the automatisation of standardised work processes. The care industry, however, struggles to implement more robots, due to the important role the human caretaker takes in the interpersonal relationship with the patient or client, which still does not allow for a standardising of most tasks without increasing the potential for harm (Hepp & Krotz, 2014).

2015)[104], which not only instrumentalise the affordance of anthropomorphism as an affect-driven induction to encourage human engagement, but encourage more persuasive bonds between humans and technologies. Social robots are consequently shaping elderly care in two ways: In allowing more data to be gathered on the elderly through the affect- or emotion-recording tracking systems with which they are equipped, and by encouraging the affect-driven bonds between human subjects and a humanoid technology that might reshape care as much as tracking will.

Next, with Chapter Seven, I begin a new discussion on machine morality streams in POT, which I see as continuous from the critical tonality on apparent moral agenthood in Chapter Five.

[104] The term 'affect economy' refers to the economical exploitation that results from fields such as Affective Programming, which do not only have an intention to understand emotions, but to transfer this knowledge into a commercial outlet. Angerer & Bösel (2015) speak of affect and psycho-technologies (in the original German text: *Affekt- und Psychotechnologien*), which are needed to detect, track, categorise, and operationalise affective states.

7. Moral Agenthood and Moral Agencies in Robots and Algorithms

Chapter summary: With Chapter Seven, I move deeper into the framework of (iii) Machine and Computer Ethics with a focus on machine morality research. I trace discussions shifting from anthropocentric agenthood to distributed agency, and from machines to algorithms equally. This chapter moves away from the projected anthropomorphised agenthood, as debated in Chapter Two, Five, and Six. Instead, I enter into new discussions on morality of algorithmic structures, which come close to my ethical and Posthumanist unfolding of tracking, but differ in the attachment to moral considerations. The concerns surfacing from this chapter, and leading to the conclusion of this thesis, are on how moral agency and accountability have become devalued and operational concepts as the operational autonomy and the distribution of agencies begin to dominate technological views on morality. The growing problem I see here is around a depreciation of human agents and decisions in the context of digital autonomy, leading to unresolved accountability questions.

On Artificial Moral Agenthood and Reductionist Morality

Summary: Wallach & Allen (2009) dominate the first section of Chapter Seven, positioning robots as 'Artificial Moral Agents' (AMAs), which points to the next defined limitation in POT research: the (2) Reductionist Morality as a limited model on the complexity of ethics as a dynamic discourse. Wallach & Allen's (2009) AMA model in robots suggests two things, in my view: Firstly, exceeding the anthropomorphic focus on perception, while pointing to the importance of the engineer and human values embedded into the algorithmic structure; and secondly, it holds onto a single agenthood position of robots and computers, while suggesting a correlation of computational autonomy and morality, which I see as problematic and reductionist to the discursive nature of morality. The second discussion in this chapter looks at Arkin's work (2009) and allows for reflection on the randomness of how morality is normed in unethical contexts of using robots.

Chapters Five and Six outlined the importance of appearance and anthropomorphism in HR and Robot Ethics and that this perspective is not seen as sufficient for a holistic ethical debate. What Chapter Six concluded with was an ethical perspective on tracking that aligns with the hypothesis of my thesis, which is that morality-led discussions in Robot Ethics are not sufficient (since too anthropomorphic) to grasp the ethical concerns around tracking or data (which require more complex agency models).

However, as I established, the ethical (Posthumanist view) on algorithmic structures might be enormously important to understand the distribution and entangled network of relations that go into code or programming, but it leaves the practical accountability questions unaddressed and does not reflect on human agent intention to sufficiently use or apply technology. As I aim to outline morality research in POT research, I increasingly establish and reflect on shifts, which I am keen to discuss in regard to their positive and negative consequences. While initially critical with the moralisation of technology, I acknowledge that the value of various morality discussions cannot be simply denied, but must be rethought, as my limitations point out.

The reason to include Wallach & Allen's (2009) work, which is situated in the machine morality camp as part of Philosophy of Technology, is that their work clearly overcomes the (1) Apparent Morality and Agenthood limitations that have occupied a substantial part of this thesis already, beginning with the anthropomorphic fallacy to examining how the companion view on robots influences the moral discussion, making it superficial and perception-based.

The focus shifts in this chapter, as I move away from social robots as bodies to zoom into their algorithmic structures (which they share with other digital systems). This is less my choice so much as a reflection on the discussions on computational morality, which simply do not pay attention to social robots. What surfaced as one interesting but confusing tendency from my literature review is that, while Robot Ethics discusses more than social robots ethically, machine morality research does not discuss social robots at all.

Social robots are partially disconnected from machine morality research; at least, they are not specifically addressed as being moral machines. In the wider research on ethical issues of robots, the issues are addressed as context-dependent studies of robots or machines, which focus on what robots are used for. In this sense, this leads to the idea that different models of robots have different kinds of moral agency concerns, even if some robot capacities are embedded in all sorts of robots throughout. Morality as a concept is hereby adapted to the aims and applications of these devices – supporting an anthropocentric view on technology.

My remark would be that social robots are somewhat treated as *step-children* in debates on moral autonomy and the morality of robots in both camps. In Robot, Machine, and Computer Ethics, as theoretically intersected fields – the dominant focus is on morality-led (less in Posthumanist) concerns and questions. The streams within differ on where and how moral agency is allocated, but it always remains an important factor. For instance, as I've already demonstrated, the moral appearance debate links to social

robots predominantly, while questions on computational capacities that might align to moral agency (assuming robots can ever have one) fit to various kinds of robots. Reflecting on the tracking and computational sophistication of social robots, I do not see any reason to exclude social robots from the wider debate on computational morality and autonomy as undertaken in POT, since their computational autonomy to monitor and to track exhibits a sufficient degree of autonomy and interaction, which allows for these to be aligned to morality research that reaches beyond moral appearances.

When it comes to (potentially) moral decisions and abilities, discussions on machine/robot morality tend to align the morally *good* or *bad* (these terms mostly remain undefined) contexts of using technology to the moral agenthood of the robot. Even if the social robot is just a *robot*, able to perform the same computational autonomy as other robots, Wallach & Allen (2009) do not consider these devices as relevant, due to their morally unproblematic context of use and the assumption that they do not have to morally judge. It could be that they indirectly always include social robots in their debates, but do not highlight them as remarkably different, or they do not consider social robots as a distinct type of robot worth reflecting upon in questions of machinic moral agency.

What this illustrates for me is a tendency in the literature of Robot Ethics and machine morality research that suggests social robots are not considered as *morally* problematic as, for instance, military robots (Lin et al., 2008). Since their agenthood is

predominantly associated with appearance, their capacities to have or develop bad or immoral intentions is neglected. I could not find any inherently technical or computational reason for the exclusion of social robots from machine morality research, since they must be able to interact, track, and behave as morally appropriate as other robots (assuming morality is a legitimate category). Hence, they are built on the same requirement as other tracking or interactive robots or computers, which are discussed as artificial moral agents, even if the levels of autonomy might vary between each type of robot.

For now, the focus that machine morality research has is on robots that are placed in morally conflicting situations, such as war or surveillance contexts, in which their agent position is crucial for making moral decisions. In current machine morality research, the focus is put on military robots or drones, in which the question is how and if robots are able to negotiate or decide on the extent of their actions or harm (Lin, 2008). Assembly robots, for example, are not morally interesting for this research, because they operate in controllable environments and do not necessarily interact autonomously with human beings, so as to give them that moral agenthood position. Whereas in environments such as elderly care and military operations, which are not inherently controllable but require 'response readiness' (Rossini, 2012) and interaction, robots could possibly harm someone.

Neglecting social robots in this sense is short-sighted, because elderly care is as much a sensitive context as military conflict, even if less *obviously* dangerous. What is

subsequently argued is that robots are worth being discussed beyond their specific applications, but in terms of their potential accountability and abilities.

Dumouchel & Damiano (2017) highlight the human fear and fascination of losing control in discussions on robot morality and autonomy. Human developers, they argue, want robots to be more autonomous, but simultaneously also fear that this autonomy will become dangerous and get out of control. Hence, roboticists are frightened to make them 'truly autonomous machines' (5). Dumouchel & Damiano (2017) refer to the struggle in which roboticists are caught up between wanting, but also *not* wanting, robots to be truly autonomous (4,5). As previously mentioned, the fear and motivation from science fiction movies and books could be influential in this debate, as much as it keeps reappearing as a driver for HR and machine morality research.

Not only are we, as humans, scared of autonomous robots, but we also aim towards changing their position in society. Robots were initially supposed to be 'workers' (Reilly, 2011: 4), which is what the word 'robot' means etymologically. The reason being that, by employing robots, these machines provide cheap labour over a long term and are not affected by human weaknesses, such as tiredness or sickness (Dumouchel & Damiano, 2017: 4). This makes robots popular in industries that need specific task-driven workers (the military, assembly lines). However, in elderly care and other increasingly growing industries where human contact is key, a robot needs to be more than a cheap and tireless worker; it must be *nice* and reliable as well. This is another reason why the moral appearance discussion cannot give any depth or substantial

ground on which to discuss morality.

While techno-philosophical researchers, such as Coeckelbergh (2009, 2010), review the moral appearance in robots as insufficient to account for moral responsibility, Wallach & Allen argue on a different level from the very beginning. They are more concerned with the *machinic* side of robots and not at all concerned with *social* robots as companions. Morality, for them, must become a machinic decision-making capacity that has no superficial affiliation at first. It must, at least, not be a question of anthropomorphic behaviour or perception, but of decision-making. Hence, they do not engage with concepts such as the 'psychopathic' robot (Coeckelbergh, 2010) and leave questions about appearance behind. In fact, anthropomorphising technology raises huge ethical problems for them, on the grounds that machines are given projected faculties they do not have, which could have harming consequences (45).

They point to two things, which I consider as valuable remarks in this discussion: The importance of the *engineer* and the values he or she embeds into the computational program and, further, the algorithmic decisions related to moral autonomy. They argue that the computational autonomy of robots enables them to evolve into 'Artificial Moral Agents' (AMAs) (Wallach & Allen, 2009; Wallach, 2010). In this sense, computers (as robots and machines) face two issues: The precise implementation of human values, and the technical sophistication to improve and adopt these values into a computational moral agenthood. Computational autonomy is the major step for them to justify why

AMAs are able to have moral 'considerations'. Wallach & Allen (2009) define morality in technology as…

> '[…]an interaction between increased autonomy and increasing sensitivity (…). With increasing autonomy comes the need for engineers to address broader safety and reliability issues. Some of those needs may involve explicit representation of ethical categories and principles, and some may not. Our guess is that engineers will add these capacities in a piecemeal fashion. Increased autonomy for (ro)bots is a process that is already well under way. The challenge for the discipline of artificial morality is how to move in the direction specified by the other axis: sensitivity to moral considerations' (34).

Their move towards the concept of moral *sensitivity* remains blurry, in my view, but this is difficult to implement artificially due to the 'immense technical difficulties [that] remain to be overcome before the robot will be able to determine, for example, which rule applies in a given situation' (189).

Wallach & Allen point out that, in their view, computers cannot yet embody or have a 'full moral agency' (26), since this is not technically possible for now. On these grounds, they further debate why and how human morality could be translated into an 'operational' or 'functional' morality that then evolves to a 'full moral agency' of the computer (2009: 26). For Wallach & Allen, there is a spectrum 'from systems that merely act within acceptable standards of behaviour to intelligent systems capable of assessing some of the morally significant aspects of their own actions' (25, 26). They continue discussing the human influence in the very making of moral rules, or 'ethical modules', which allow for the machine to establish its own set of rules in a Utilitarian ethical manner, upon which to build their moral actions (Dumouchel & Damiano, 2017:

175).[105] This requires that these devices must be autonomous enough to execute and judge within the restrictions given to them by the developer (Dumouchel & Damiano, 2017: 175), while the moral and universal rules must be clearly defined as precise instructions.

The engineer is crucial in this process, according to Wallach (2010) who asks; 'Whose morality or what kind of morality should be implemented in ro(bots)?' (243). Because of the importance of the human developer in implementing moral judgement into the machine's programming, Wallach & Allen suggest and urge for the development of 'Engineer Ethics' (Wallach & Allen, 2009: 25). The difference between what they highlight and what Chapter Six discussed with tracking being an ethical process is that they understand the values and concepts are being implemented in a *linear* and stable causality into a *neutral* technology. The embedding process of the human values into the computational is not questioned by them as being, in itself, a mediating or ethically entangled process, but is described as a linear path between the human instructions and the algorithmic structure executing the rules.

Wallach & Allen state that, therefore, the developer must be able to set the rules clearly and precisely. For me, this not only implies that morality is a rule-based system that can be implemented, but that it also requires the engineer to have great overview in being able to anticipate the consequences emerging from the implementation of these rules

---

[105] Dumouchel & Damiano (2017) write on the Utilitarian ethical tradition: 'Since it has long been common practice in various versions of utilitarian doctrine to quantify the moral value of different options for the purpose of comparing them, utilitarianism plainly recommends itself to anyone seeking to thoroughly "mechanise" moral reasoning' (231).

and on the possible behaviours of the AMA that, by executing these rules, becomes 'ethical' (16). They consider that this new position to which machines are elevated requires their moral accountability for the outcomes, as much as for the actions that these machines do not yet have. Hence, for them, machines still lack the sufficient computational autonomy necessary to reach this state of an AMA, but it does not mean that they always will. Wallach & Allen suggest that developers can support the development of the machine's moral considerations by establishing an…

> '[…]open-ended system that gathers information, attempts to predict the consequences of its actions, and customizes a response to the challenge. Such a system may even have the potential to surprise its programmers with apparently novel or creative solutions to ethical challenges. Perhaps even the most sophisticated AMAs will never really be moral agents in the same sense that human beings are moral agents' (16).

The first concern I identify in this AMA concept is that Wallach & Allen consider the embedding of human moral values as a *neutral* and linear process into a *neutral* technology. This position originates from an instrumental view on the computer as an extension of the human developer's values, and has only one positive aspect in this case: it allows for moral accountability to be positioned clearly with the human agent. Gunkel (2016), however, points to the limitations of thinking in an instrumental perspective by saying that:

> '[T]he instrumental theory of technology, which had effectively tethered machine action to human agency, no longer adequately applies to mechanisms that have been deliberately designed to operate and exhibit some form, no matter how rudimentary, of independent action or autonomous decision-making (241).

But, as they have argued themselves, machines are increasingly complicated devices and systems that, on the one hand, are supposed to judge morally but, on the other hand, are considered tools. This creates an oxymoron in their model of the AMA, which is not that uncommon in this research camp and follows a cybernetic perspective on technology in which attributes can be simply *added* or *extended*, where the dynamic in this process is insufficiently reflected upon (compared to media theoretical or Posthumanist research).

On the one hand, the machine is only a neutral instrument; on the other, it should be an artificial moral agent. The two angles – between machines being neutral instruments or highly developed systems – creates a conflict that the philosopher Zylinska (2014) addresses. She writes that the 'anthropocentric moralism (where values are being laid out without questioning the process of their fabrication and the conflict in which they always exist with some other values)' (72) is joined by us encountering 'the danger of falling prey either to anthropocentric moralism or to delegating authority to technology which remains underpinned by instrumentalist assumptions' (72).

Beside supporting this conflicting view on technology, another issue that Wallach & Allen's (2009) concept raises is the view of the robot as an 'individual actor' (190) and not as 'moments of complex technological system' (191), which, according to Dumouchel & Damiano (2017), is a problem. This will be picked up in the next section of Chapter Seven on Floridi's (2014) work, which overcomes this 'anthropocentric conception of agenthood' (187) and suggests a different conceptuality towards

distributed moral agencies. A further issue I have with Wallach & Allen's work is that they see the values of the engineer as universal, agreeable, and stable norms. Even if they urge for an ethical acknowledgement of these having to be considered, they do not unpack who decides on the making of these norms and on what grounds.

Arkin's (2009) research on military robots shows how the norming of machine rules becomes a dilemma if there is no clarity on what capacities robots need to be able to judge and understand the context of their application. His work amplifies why the process of morality conceptualisation is more than finding agreement on universal moral rules to be executed by a machine. The expectation to assign clearly defined universal values on good and bad action has already been mentioned as one major concern, which the Posthumanist Braidotti (2006) declares as a 'hindrance' (15) of moral philosophies.

It shows that Arkin struggles at this very point, but he resolves the struggle differently, by appropriating what moral *goodness* means to the context in which it must be applied. Arkin's research amplifies how computational autonomy gets increasingly and problematically synchronised with moral autonomy, while making it gradually impossible to allocate agenthood or accountability within these abstract *moral* decision processes. The definitions of *good* and *bad* behaving robots are not only flawed when expecting the universality of these terms, but becomes even more flawed if these terms get adapted to the context of use and do not provide any stable norming framework.

Arkin's (2009) research addresses the dilemma around military robots as autonomous and, therefore, morally responsible agents. The major issue being; what is considered to be a *good* robot in the context of having to do *bad* things is highly debateable. This shows that morality is, in the end, adapted to the context of needs and tasks, which makes the idea of a universal norming problematic. It asks if the adaptation of morality to the need of a context has not simply become the making of random rules and labelling them as moral. The moral reference point beyond, for instance, aiming for more efficiency or accuracy in rule-based decisions seems to disappear if morality has adapted that flexibly. This thesis argues that by adjusting morality to a context and need, as to war or conflict situations in which good and bad depend on what side the robot fights, it becomes absurd.

Arkin is aware of this trap, but suggests a creative and, in my view, unrealistic repurposing of robots to keep their moral value nonetheless. The robot could take over the position as the moral guide in stressful and conflicting contexts to prevent humans from becoming immoral agents. According to Arkin (2009), the moral intention programmed into military robots is not meant for them to be morally good agents, per se, but for robots to be able to minimise the risk of civilian death or to avoid exaggerating a conflict beyond what is strategically necessary (according to the developer) (30). What Arkin concludes from his research into U.S. Army robots is 'that robots not only can be better than soldiers in conducting warfare in certain circumstances, but they also can be more humane in the battlefield than humans' (30). He claims that in an ideal (moral developer's) world, the robot would basically prevent

the human from becoming immorally carried away to do more harm. What he wishes for in this context is that if 'weaponized autonomous systems appear on the battlefield, they should help to ensure (...) humanity, proportionality, responsibility, and relative safety' (33).[106]

In the research on morality in military robots and their moral accountability, unanswered questions reappear such as: What is moral? Who is the agent? And, can computational autonomy be held morally accountable for the execution of rules? Since what the robot is faced with, in this position, is a conflict between its autonomy to execute precise algorithmic rules, and having to regulate the immoral human agent.

Ideally, Wallach & Allen (2009) are correct in arguing that machines are going to add 'novel or creative solutions to ethical challenges' (16), but this thesis struggles at this point to agree with the linearity and instrumental view on machines, the reduction of morality to algorithmic rules, and the *creative* adaptation of moral norms to contexts. As such, the guidance that morality gives in being a purposeful and reflective concept appears much more flexible and adaptive than might be useful when answering for uncomfortable moral questions.

---

[106] Sharkey (2012) critiques Arkin's conclusions by arguing that 'Arkin's anthropomorphism in saying, for example, that robots would be more humane than humans does not serve his cause well. To be humane is, by definition, to be characterized by kindness, mercy, and sympathy, or to be marked by an emphasis on humanistic values and concerns. These are all human attributes that are not appropriate in a discussion of software for controlling mechanical devices' (219). I am critical with both assumptions; on the robot being able to be humane, as much as with saying that human individuals are, per se, kind or merciful.

Debating morality, moral agenthood, and moral accountability has not only created multiple conflicts of interests between the warring parties, but also aligns research to ethical boundaries beyond doing *good*. What moral *goodness* means, then, is only bound to context-specific requirements that allow for robots to evolve into moral agents if their behaviour conforms to what is programmed into their complex (man-made) set of computational rules (Dumouchel & Damiano, 2017: 187). This leads to a reductionist morality being simplified to a practice of accuracy and algorithmic rules, and leads to a conceptual framing problem that Wallach (2010) noticed as well. He claims that:

> 'Maximizing goodness or achieving justice can be rather vague ends when the agent needs to discover what the goal means rather than having the end accompanied by a top-down definition' (247).

Hence, what the work of Wallach & Allen and Arkin showed was that questions on moral universalism or reductionism have complicated their own discourses in which these were supposed to assign moral *goodness*.

On the one hand, the autonomy of computational algorithms complicated the accountability question, because robots or machines must be able to interpret and execute non-neutral human instructions, but are not yet accountable agents. On the other hand, as this thesis highlights critically, the precise execution of specific rules cannot be taken as sufficient to become morally accountable or to regulate human accountability, as Arkin would suggest, due to the lack of technical sophistication and moral reasoning of machines or algorithms.

This section has allowed for two main things to emerge. It successfully overcame the perceptive companionship angle of robots by deepening the agenthood discussion as one situated in the machinic and computational dimensions of robots. Furthermore, while the AMA concept of Wallach & Allen (2009) goes beyond the perceptive level of agenthood and morality, it does not go beyond moral reductionism and a stable and singular agent position in the machine. They have further limited the conceptualising of morality to a rule-based view on computational *moral* decisions, which are operational algorithmic decisions. This leads to an additional challenge in which the human engineer is responsible for embedding moral values into the AMA, but then must also assume that this 'system may even have the potential to surprise its programmers with apparently novel or creative solutions to ethical challenges' (16) and present its own moral decisions one day.

If the AMA gains sufficient autonomy (it is not clear when and if this will happen according to Wallach & Allen), it might then be used to support the human agent to regulate the scope of immoral behaviour in moments of stress or conflict, as Arkin (2009) suggests. However, in none of these scenarios can the computer, as a device, be held accountable for any of its behaviour or consequences. Nor has it been clarified on what requirements or parameters an artificial computer agent is to be identified as an agent. What appears also problematic is that the scenarios outlined are prolific and supportive of highly reductionist and rules-based models of morality, orientating

themselves around a mixture of Utilitarian and Kantian ethics[107], by aligning algorithmic decisions with moral actions to project moral agenthood.

[107] Wallach (2010) addresses the still unresolved intersection of will and autonomy in Kantian ethics. 'Kant's contention that *will* and *autonomy* are necessary for an entity to be a moral agent. The ability to function as an autonomous being, or the capacity to will, suggest faculties beyond pure reason. However, little is understood regarding the manner in which Kantian will and autonomy are supported by and emerge from the capacity to reason and other cognitive mechanisms' (246). Wallach also points to the importance of unconscious or emotional decisions in human morality, which cannot be translated into computational code, but still influence moral decisions. Hence, the research on AMAs must look beyond cognitive mechanisms of moral decision-making (245).

Towards Distributed and 'Mindless' Morality

Summary: The second section in Chapter Seven reflects on two further models with which to discuss moral agency, which represent a shift in POT, from understanding machines as single agents to understanding machines as digital system agencies. I continue the already elaborated aspects from Chapter Six on tracking, where I pointed out why the computational autonomy of robotic machines is an important level on which to discuss moral autonomy and agency. Brey (2014) proposes thinking in terms of 'ethical structures', while Floridi (2014) suggests reconsidering the 'anthropocentric agenthood' (187). In contrast to Wallach & Allen, Brey and Floridi overcome the singular agenthood and distribute moral actions and moral outcomes into the algorithmic structure. This section will conclude with two examples of algorithmic autonomy, which illustrate that, practically, the question on how to allocate moral actions and factors within semi-autonomous algorithms is not only difficult, but it is impossible to allocate any moral accountability to such complex technology.

So far, I have moved beyond the appearance-based model of technological agency and will, next, shift the view on digital technology as a single agent, which proved to be a tempting, yet superficial, way to view robots (because of their humanoid bodies). Even if robots are not their research focus, I examine Brey's and Floridi's morality and ethics research in detail to examine this shift. They suggest distributing moral agency into algorithmic agencies and discussing the wider network of agencies around moral outcomes and moral actions; not to assign moral agenthood to machines, as Wallach & Allen suggested. The shift they suggest in distributing morality favours a better computational understanding of algorithms by adapting morality to an algorithmic understanding. The issue that emerges from their work is that a distributed model of morality diminishes the accountable agent and removes the question on any practical accountability of technology. While this allows for a better understanding of algorithms, it devalues morality as a concept that might be able to address the consequences and responsibilities of agent behaviour.

These two concepts are already highly complex attempts to understand ethical discussions aligned with computational autonomy. Both consider the increasing computational autonomy as influential in the context of moral accountability. They reflect on why a moral accountability cannot be stabilised or attributed to algorithmic structures after all; even if morality increasingly dissolves into a theoretical and abstract model, accountability is removed from these discussions by being assigned to the human agent only. Brey and Floridi are interested in a flattened, almost Posthumanist

view on morality, as a horizontally laid out relationship and less as a hierarchical evaluation of anthropocentric agenthood.

Brey (2014) suggests thinking of 'ethical structures' (135), which at first sounds as if he has liberated this view from a limiting intention to moralise technology anthropocentrically into good and bad behaviour. For him (2014), structural ethics 'are social and material arrangements as well as components of such arrangements, such as artefacts and human agents' (135). His model…

> '[…]has three aims: (1) to analyze the production of moral outcomes or consequences in existing arrangements and the role of different elements in this process; (2) to evaluate the moral goodness or appropriateness of existing arrangements and elements in them, and (3) to normatively prescribe morally desirable arrangements or restructurings of existing arrangements' (135).

Discussing technology ethically instead of morally was already raised as an alternative view in Chapter Three when reflecting on the limits of moral philosophy. But even if the ethical perspective on socio-technological agencies allowed to understand the entanglement of agencies and, later, allowed for an ethical unpacking of tracking as a network, this view has similar limitations on accountability questions.

However, the ethical framework does not intend to advocate for a stable or accountable agenthood in technology, neither in the human agent, since it is positioning its exploration differently to moral discussions, trying to overcome the hierarchy of agenthood. Brey seems to have found a middle ground in thinking technology ethically without fully removing the agent. He discusses morality as a distributed system of

social and material networks and arrangements, not only as a 'residing' (137) concept in one agent. This seems very appropriate to grasp computational structures as well, even if his work speaks more broadly about technology and artefacts (but also about CCTV cameras).

The important shift in Brey's research, compared to the previous work of Wallach & Allen, is that he is interested in a structural and not an *agent*-focussed conceptuality of morality; what he describes as a shift from individual ethics to structural ethics. He, therefore, returns to the terminology of ethics as aligned to what Chapter Three suggested, to be able to look at the structural unfolding rather than the moral evaluation. Within this approach, Brey argues, morality can be reflected upon in different ways; for instance, by looking at the 'moral factors' (138) influencing it, which can be assigned as moral outcomes or moral behaviour. For Brey, this ethical focus allows us to reflect on various factors that arise in thinking about technology ethically. He points, for instance, to norms in society that shape our way of thinking and behaviour, but also the roles that artefacts have in society (139).

However, Brey's focus on 'moral outcomes' in thinking technology becomes problematic, because of his suggestion that these networks own a 'moral goodness' attached to them, which, again, moralises technology as good or bad while not saying why this is useful. Yet, what can be agreed is that Brey is aware of and defends the view that technology is neither morally neutral nor can it be made responsible. He sees structural ethics as a network of moral factors, which are outcome-oriented or

behaviour-oriented and, therefore, allow for an understanding of technology in a moral way, but in relation to behaviours or outcomes, not in being an accountable agent (138).

In Brey's work, the process of trying to include moral accountability into this scheme becomes slightly abstract, in my view. For him, moral accountability could *theoretically* be given to different agents in this ethical network (independent of these being human or robot entities), but, in case none of these agents can be made accountable, only 'human agents bear responsibility' (138). This is one way of saying that *practically,* only human agents bear accountability. However, the ethical structure is his way of suggesting that technology or machines can have moral agency as an ethical structure, but no moral agenthood as agents.

Yet, what would this mean in reference to a tracking algorithm and its 'moral goodness'? Such could be reviewed by looking for the moral outcome, but not expecting any moral accountable position within its digital system. Brey continues to debate digital systems in anthropocentric terms of having good or bad moral outcomes, which I consider problematic and challenging to some extent, since I question the ability to pinpoint any goodness or badness within the algorithmic system. Nonetheless, I do acknowledge that, in the wider context of how these systems effect human lives, his approach is important, despite the moralising tendencies.

Floridi (2014), on the other hand, joins this discussion with the concept of a 'mindless morality' as a strategy to overcome the fixation on a single agenthood discussion. Despite proposing an interesting and complex perspective on algorithmic morality, his

model only reinforces the concern with (2) Distributed Morality and distributed moral agency, as I will explain next. His suggestion of thinking through a 'distributed moral responsibility' (DMR) or 'mindless morality' (188) within algorithms considers the moral consequences of technologies beyond agent intentionality. Floridi comes into the discussion on machine morality with a clear intention to shift the focus of ethical questions in computer technologies (and robots) by forming Data Ethics (Information) Ethics, as a new, challenging field in which ethical consequences of computational technologies are debated (Floridi & Taddeo, 2016). Floridi is as interested as Brey in an ethical and structural view, as his critique on the anthropocentric agenthood implies. His first suggestion is to overcome the anthropocentric agenthood. He states that:

> 'An entity is still considered a moral agent if (i) it is an individual agent, (ii) it is human-based, in the sense that it is either human or at least reducible to an identifiable aggregation of human beings, who remain responsible as the only morally responsible source of action, like ghosts in a legal machine' (Floridi, 2014: 187).

This partially aligns with Posthumanist ethics, even if their focus would not approve the *leftovers* of moral evaluations that Floridi embeds in his thinking. While Floridi claims that there is no ethical thinking beyond intentionality, for him, the agent is not required to exhibit such exclusively. This indicates that Floridi and the Posthumanist debates might even share a similar ambition to leave the anthropocentric agenthood angle behind, but Floridi continues to moralise technological actions by not being clear what moral concept he applies to their evaluation (while he is very clear in critiquing individualist moral philosophy as inappropriate to understand digital information structures).

Floridi critiques the anthropocentric agenthood focus by arguing that the traditional angle on moral agents demands that the agent is morally good within what is defined as the 'moral threshold' (188). He outlines a different path to acknowledge that the computational autonomy in machines has complicated moral agencies. Floridi is primarily interested in software and algorithms in Information and Communication Technologies (ICTs), compared to Brey, who refers to technology in a wider sense and, particularly, to CCTV cameras. Floridi suggests to redefine the limits in thinking of 'moral agents' in technology by liberating this position from the focus on an anthropocentric agenthood. This has a huge advantage, since, by overcoming the 'anthropocentric conception of agenthood' (187), the complexity of algorithms can be understood much better. However, by doing this, the disadvantage is that the distribution of agenthood into agencies makes an allocation of an accountable agent impossible.

Floridi goes even further and detaches accountability from moral responsibility, which is a most problematic point for me, considering the operationality of moral responsibility seems to devalue any reason for keeping moral questions attached to technological systems. This formulates my major limitation on the (2) Distributed Morality model, besides it having removed single agent thinking. Floridi's model for thinking morality is built on an abstract and complicated scheme, in which he discusses different 'Levels of Abstraction' (LoA) of informational structures.[108] The LoA model

[108] 'The Method of Abstraction comes from modelling in science, where the variables in the model correspond to observables in reality, all others being abstracted. The terminology has been influenced by

is, according to Floridi (2014; Floridi & Taddeo, 2016), a way to look at moral agents beyond intentionality, moral states, or the moral nature of the agent, and to think of 'mindless' morality. Floridi hereby closes the circle between moral autonomy and computational autonomy.

For Floridi, the allocation of moral responsibility is possible within technology, since responsibility can be looked at in an operational way. This means finding the source of the action, like an error in the algorithm, and assigning the *responsibility* to the position in the algorithm that creates the error. Consequently, the source of error becomes 'morally answerable' (2016: 6). The important thing for Floridi is that this model can address erring agents, which can learn from their *wrong* behaviour. What is required is for them to be autonomous enough to learn and to be able to change the rules in the system to improve their behaviour (7).

The major reason for him moving responsibility away from accountability is to highlight that computational systems are never fully accountable for their actions, but that, within their structure, there is a responsible agent whose behaviour and action can be adjusted (7). Even though Floridi denies and critiques the anthropocentric fallacy himself, he strangely continues to call these 'responsible agents' owning 'good' or 'evil' (7) intentions.

---

an area of Computer Science, called Formal Methods, in which discrete mathematics is used to specify and analyse the behaviour of information systems. Despite that heritage, the idea is not at all technical and, for the purposes of this chapter, no mathematics is required' (Floridi, 2014: 190).

His model illustrates a few problematic points for me, not only because somehow Floridi seems to again return to single *agents* despite his very critique of these, but because he has detached responsibility from accountability, while also detaching accountability from agenthood. I consider this step highly problematic, due to the transformation of moral responsibility to an agent-less operational causality. According to Eshleman's (2014) and Noorman's (2018) description of moral responsibility, this means removing the direct association between moral responsibility and a potential punishment or blame as a consequence of immoral or harmful irresponsibility. Interestingly, a critique on Floridi's approach comes exactly from Brey (2014), who reviews Floridi's solution as 'unsatisfactory because moral agency has traditionally been identified strongly with moral responsibility' (140).

Miller (2010), one of the few who emphasises the role of the human agent in POT discussions, points to the relevance of human moral responsibility and outcomes in saying that:

> 'The relevant human players, systems designers, and software engineers, for example, and not the computers, have collective moral responsibility for any epistemic outcomes' (Miller in van den Hoven, 2010: 5).

I aim to highlight the practical issues in assigning accountability to algorithms in a next step, to illustrate why Floridi's approach might be conceptually interesting, but is practically distancing POT views on technology from the practicalities of applied ethics. To do this, I look at two algorithmic examples that illustrate how complicated it has become to understand moral intention on the one hand, and agenthood on the other.

Also, how dangerous it might be to think of morality only in operational terms. The examples will also show that the distributed thinking of morality is useful to understand computational agencies, but is not very favourable to address any accountability beyond the human agent developing or programming software, even if such an agent seems to have lost full agency on the ethical consequences.

On Amoral Algorithms

Summary: The next discussion circles around two examples of algorithms, which exhibit new autonomies, since the increasing computational autonomy has complicated the discussion on algorithmic morality. Even if the shift towards agency distribution is argued as a valid and enriching step previously, it has complicated the practical confrontation with, and the consequences of, algorithms. This is especially problematic and potentially unethical, when algorithmic systems lead to unintentional but severe harm of human users, exhibiting how neither has their autonomy, nor has their accountability been sufficiently understood.

The first example of algorithmic complexity that escapes a moral agenthood and clear moral intentions is drawn from Wallach & Allen's (2009) work. I do not hereby return to their work or views, but just make use of one example in which they have identified concerns with computational autonomy; one I find appropriate to make my point. What they illustrate is how the increasing computational autonomy in algorithms has complicated the search for morally accountable agents, while they were still taking about single agents. Even if I considered this approach flawed and reductionist to fully understand computers or algorithms as networks, it offers a simply way to allocate moral action/decision to one point/device/moment, even if this remains a theoretical allocation only.

Wallach & Allen (2009) indirectly anticipate that considering computers as agents is insufficient (even if this goes against their AMA model), as they unpack an ethical dilemma concerned with the lack of accountability emerging from malfunctioning algorithms in medical machines. Them quoting the research director at Google, Peter Norvig, shows the severe consequences for human lives from, in my view, amoral but highly ethically influential medical software errors. Norvig highlights:

> 'These are errors like giving the wrong drug, computing the wrong doseage, 100 to 200 deaths per day. I'm not sure exactly how many of those you want to attribute to computer error, but it's some proportion of them. It's safe to say that every two or three months we have the equivalent of a 9/11 in numbers of deaths due to computer error and medical processes' (Norvig, quoted in Wallach & Allen, 2010: 22).

Wallach & Allen do not bring this example up to question their AMAs, unfortunately, but do advocate that 'the harms caused by today's (ro)bots can be attributed to faulty

components or bad design' (Wallach & Allen, 2010: 22). Their quote reads as if design faults emerge from the technology having a bad intention to malfunction or the engineer having an immoral intention. They do not discuss what it means practically if harmful and ethically severe consequences occur *despite* there being no intention to do harm without any identifiable malfunction, which is what the second example will illustrate.

Further questions arise: Who/what is accountable for a bad design? Could the 'Artificial Moral Agent' (AMA) ever be held accountable? What would be the threshold between an artificial computational agent not being able to execute the rules implemented into its systems, and the human developer not providing it with the correct parameters to do so? Machine and Computer Ethics continues to look for the *moral moments* in these systems, but does not resolve any moral concerns or outcomes sufficiently. Further, the gap between the theoretical complexity and the practical insolvability seems to grow.

As Brey and Floridi suggested, shifting the discussion from *agents* to *agency,* in order to negotiate accountability and autonomy as an ethical network between the human and the machine, does acknowledge the complexity of algorithmic agencies, but does not address questions on practical agent accountability, which only amplifies the gap I see between theory and application, between meta ethics and applied ethics. If looking at the medical algorithm malfunction through Floridi's morality model, I cannot find or allocate what Floridi calls the 'evil' error in the algorithms that could be held even operationally responsible, despite never being accountable. With Brey's ethical model, I might be able to identify the *bad* 'moral outcome' that the algorithmic ethical structure

brings with its application, but what if it was programmed to lead to this outcome? With both, the question on the system's inherent moral judgement and on moral decisions seems unresolvable.

In my view, the medical algorithm has no intention to lead to negative consequences, which are only negative for the humans, not for the algorithm, since the algorithm executes what it is made to do, even if this includes an error in the system. Such an error is seen as *bad* or immoral only according to what it leads to; not because it can be understood as an immoral decision, as Wallach & Allen would suggest, but, instead, it proves to be a moral factor influencing an ethically problematic outcome of the ethical structure, as Brey would argue. This implies that the ethical outcomes could be the most important factors from which moral questions must be addressed when discussing morality in technology. Therefore, the only way to think of the ethical consequences is when allowing for the algorithmic structure to unfold such autonomy and complexity, and not when harms happens, and to then retrospectively untangle the network of agencies and search for the *one* accountable agent or to assign blame by projecting expectations into 'mindless' systems. Assigning a moral intention retrospectively or anthropocentrically into these systems seems absurd, and practically useless, in my opinion.

Gunkel (2016) displays a different case of an, in my view, amoral, but equally consequential, autonomous algorithm. Instead of looking at a specific programming error in the system, he looks at banking or financial algorithms and amplifies how

autonomous learning machines have become morally uncontrollable. Gunkel holds on to Winner's points from his book on *Autonomous Technology* (1977), published more than 40 years ago, to reinforce that:

> '"To be autonomous," Winner (1977) argues, "is to be self-governing, independent, not ruled by an external law of force" (16). The phrase "autonomous technology," therefore, refers to technical devices that directly contravene the instrumental theory by deliberately contesting and relocating the assignment of agency. Such mechanisms are not mere tools to be directed and used by human users according to their will but occupy, in one way or another, the place of an independent and self-governing agent' (2007: 239).

However, the operational autonomy has drastically improved, even if what self-governed means should be rethought. According to Gunkel, financial algorithms can be understood as *learning* systems, which become autonomous to some extent. This does complicate the question on their increased agency and decision-making, in suggesting that, despite this, they cannot be held accountable. These mechanisms are 'designed not only to make decisions and take real-world actions with little or no human direction or oversight but also programmed to be able to modify their own rules of behaviour based on results from such operations' (239).

Gunkel refers to the problem of allocating intention and accountability by looking back at the worldwide financial market crash that originated in the U.S. in 2010 (see next quote). He reflects on the time when human traders used to control stock markets completely, but that from the 1990s onwards, this agency was increasingly transferred into intransparent algorithmic agencies. These new models of regulation were faster in trading and comprehension, learning from and adapting to unexpected opportunities

more quickly than humans. Gunkel (2016) cites Patterson (2012) to point out that this led to 70 per cent of international trade being generated by 'autonomous' machines beyond the influence of human agency. These machines organised exchanges from mortgage payments to retirement savings (239). Gunkel (2016) then points out to where this shift in agencies and power dynamics ultimately led. He writes:

> '[T]he unanticipated social consequences of this can be seen in a remarkable event called the Flash Crash. At about 2:45 p.m. on May 6, 2010, the Dow Jones Industrial Average lost over 1,000 points in a matter of seconds and then rebounded almost as quickly. The drop, which amounted to about 9% of the market's total value or $ 1 trillion, was caused by a couple of trading algorithms interacting with and responding to each other (241).

In this scenario, no single human agent nor any 'bad design' can be held accountable for the kinds of consequences that emerge from the complex autonomy of banking algorithms (which are nonetheless managed my human agents). The question in this case is less; if human agents could have stepped in to stop a financial crisis, but what financial systems and capital groups prevented this from happening. It is unlikely that an algorithmic system is that powerful (technically speaking) so that human agents have no way of controlling it, and yet, the autonomy these information structures, which are embedded in other power structures, operate vast amounts of information in real time, by also managing endless decisions. This cannot be underestimated in ethical terms.[109]

---

[109] The influence of algorithms on our lives should not be underestimated, as Gunkel (2009) warns. These systems do have agency that shows, for instance, in recommendations from services such as Amazon, Netflix, and Google search results. These might be mistaken as information providers or suggestions at first, but they do influence the relationships to knowledge, culture, information, and what kind of people

The examples emphasised that, while the increasing algorithmic autonomy of digital system is encouraged despite having severe consequences on human lives, the human agents are weakened in their accountable position, but are the only accountable agents. Further, the complexity and lack of transparency in algorithmic actions simply escapes the allocation of any moral intention or error, because of its distributed structure. I have already suggested that the ethical negotiation of power structures and discourses around robots, as a technology of algorithmic structures, must be understood more holistically, instead of supporting anthropocentric or anthropomorphic models of moral reductionism. However, I must point out as well that the position of the accountable agent, as found in simpler and instrumental technology models, has been devalued or remains undebated because of this. Again, the previous POT streams (including their limitations) are not totally dismissed here, since they exhibit valid theoretical insights, but they seem unable to suggest practical solutions to the concerns they address. It seems to me that these start off as discussions in applied ethics, but increasingly transform into meta-ethical streams.

What algorithmic structures have illustrated to me is how complex it has become to negotiate the human input in relation to the technological capacities and processing power. However, I stressed that *the* algorithm is not an entity and does hold an accountable fixation point for its technological capacities and processing abilities, even if these lead to ethical consequences. Defining what these are and who is in charge must

---

we will meet or know about, consequently. (241) For me, this is a good example on computational technology being inherently ethical when negotiating power structures and relationships.

be put up for further discussion since, for now, only human agents can be held accountable dealing with any outcomes, as they are the only agents affected by these and able to evaluate their outcomes.[110]

To conclude Chapter Seven, I pointed out that certain limitations of the (2) Reductionist Morality and the (3) Distributed Morality models only reinforce a reductionist or a distributed model of technological agencies, while aiming increasingly towards a Posthumanist ethics without addressing these shifts and their consequences sufficiently. In my view, these consequences ground in the uncritical and positivist tendency in (iii) Machine and Computer Ethics and wider morality research, which is critical towards the anthropomorphic fallacy in (i) Robot Ethics, but creates new concerns as morality is reduced and simplified to numeric and algorithmic rules. In a wider sense, these views have transformed morality research from understanding technology *morally* towards transforming the concept of morality *algorithmically,* and they have increasingly (but more subtly) undercut questions on accountable agents.

I am concluding this thesis by reflecting on my findings and by suggesting further research.

---

[110] At this point, the question of legal accountability should not be underestimated and might even reflect on how much the machine is considered accountable for a negative outcome, or the user, developer, buyer etc. These questions have huge practical implications that could affect elderly care as well. Hence, debating agency and accountability of machines is not a theoretical endeavour - as I highlight multiple times, but cannot unpack in detail.

Conclusion

Dumouchel & Damiano may be right by saying that we are, in fact, afraid and equally fascinated by robots. This could be one reason why roboticists or programmers try to equip robots with morality *buttons* as they seek to make these devices more autonomous but, also, nicer. As robots gain autonomy, many expectations are projected onto them and some distort the role of the developer's responsibility and agency as much as the norming of the (imaginary) morality buttons. Simultaneously, the fear that robotic autonomy will become dangerous, get out of control, and become an evil machine is countered by views of the caring companion; the good companion robot.

Before concluding, I want to revisit my introduction briefly and the two identified conflicts, which I saw around the ethics of social robots used in elderly care. The first conflict was considered to be on hybrid and conflicting agency models of social robots, which I saw as either being anthropomorphised companions or as data tracking devices. My intention was to find out why the latter perspective is neglected in Robot Ethics and what supports or rejects this neglect. I saw the second conflict situated in the conflation between having either moral views or ethical views on robots, which relate to how we project values into them as a digital technology. Analysing these conflicts – between agency models and ethical perspectives – enabled me to form certain themes around agency, ethics, and morality and discuss these in three wider frameworks. Next, I unpack my findings less chronologically, but, rather, as thematically clustered.

On the Companion Position and Anthropomorphic Morality

I addressed the anthropomorphic view on robots by critiquing its superficial perspective on agency and its limiting status as anthropomorphic ethics. I identified that Robot Ethics discusses moral agency of social robots through an (1) appearance-driven morality model, which originates in HR. There, social robots are intensely positioned and promoted as companions and not as machines, so that implementing these technologies into new care or heath contexts, such as elderly care, becomes easier. This view is mistaking the robot to be two things it is not; a single entity and a human-like moral agent. I showed that the superficial resemblance to human bodies is insufficient as a base to discuss actual agency models.

New and progressive agency models on technology are better understood from an 'ethico-onto-epistemological' agency model, as Barad would suggest, but not from a perceptive or moralising view on technology as a separate entity. What I specified was that the ethical agency of robots must not be grounded (only) on its humanoid shaped design, nor on its simulation or the extension of human values or agency, since these inference aims are instrumental and underestimate technological agencies in their holistic unfolding or becoming. A robot is a non-human agency entangled with human agencies, leading to new ethical concerns, such as being a possible dataveillance structure. The Posthumanist view pointed to the practical entanglements and allowed me to understand data collection as a holistic process. Yet, at the same time, this

perspective did not enable me to reflect on applied concerns around the practical use of robots.

However, I do not conclude that the research on the psychology or the perception of robots should be underestimated or left out from ethical canons, even if this angle is important for Robot Ethics, because the human affectedness through anthropomorphism does influence the Human-Robot-Interaction (HRI) immensely. However, the acceptance and interaction with technologies such as robots does occupy various disciplines, and I view these canons as rhetorically imprecise when discussing expectations or capacities of robots. I highlighted this by repeating how terms such as good, natural, social, bad, or even evil are attributed as mostly rhetorical to project qualities onto robots, but do not capture their actual capacities.

I stated that Robot Ethics must rethink its mind-set on what it is aiming for when trying to simulate humans through robots, especially because morality is not made; it is debated as a reflexive concept. Robot Ethics operates with a limited framework on thinking ethics, one that conflates ethical concerns with moral implications and deliberates on morality with an anthropomorphic and appearance-based approach to agency. This confirmed my assumptions on how the agency position of robots depends closely on the moral or ethical model it supports and vice versa.

<u>On Moral Concerns around (Social) Robots and their Limitations</u>

As I struggled with the superficial view around robot agenthood, I further highlighted ambiguous tendencies in POT discussions, which move beyond anthropomorphic and

superficial agent models, but remain stuck in a moralisation of technology. I identified the difference between anthropomorphic and anthropocentric morality discussions, which are both morality driven and built on human expectations of robots. Since the anthropomorphic view was at first my central theme in the ethical discussion on social robots, I traced it back to its origins in Humanoid Robotics (HR), through which I showed that understanding social robot anthropomorphically conflates their humanoid appearance with their technological capacities as tracking devices. With this discussion, I drew attention to the importance of computational interaction in understanding the technological capacities of a robot, beyond anthropomorphic projections, and to expand the ethical discussion.

As moved on, I identified tendencies of reductionism and expectations in the anthropocentric camp as well, which I will pick up in more detail towards the end of this conclusion. Overall, what I determine from both perspectives is that morality-centred and human-centred focus on robot agency create a legitimate discussion to some degree, but can be limiting for a holistic ethics on robots in all their agency facets; especially since drawing norms on robot behaviour from human behaviour still proved to be a projective and schematic approach to grasp robotic agency. The danger I saw in this context is to mistake robots as neutral tools or to view their agency as an extension of human agency only. Hence, I urged to take robots much more seriously; not because these are deceptive humanoid agents or potentially bad ones, but because these are complex agencies, *despite* not being human agents.

On Robots being an Inherently Ethical Technology

Social robots raised concerns from various ethical angles as I advocated to understand them – and digital technology in a wider sense – as inherently ethical relations and as an inherent ethical network of human and non-human agencies. By looking at these devices as ethically intertwined agencies made of, and influenced by, human expectations, agency models, values, data sets, privacy, and autonomy epistemologies, I argued that the production and use of robots is no one-way street. It is a process that reciprocally influences human relations, agencies, and values, and feeds back into new sociability models, which do not have to be deceptive, but are nonetheless not the holistic, ethical dimension worth unfolding. This conviction emerged as I suggested thinking of ethical networks and entanglements of norms, values, and relationships in technologies, which are influenced by the humans creating them, but which are not an extension only.

I showed that a robotic device and technology expresses and embodies human values and thoughts that went into its making, but it is not simply an extension of these. My extensive analysis on tracking illustrated this entanglement in detail, highlighting that technological capacities and human values become more than the sum of their parts and complicate a reversed detangling of agent positions. By doing this, I also elaborated on why understanding tracking better requires looking at disciplinary values and goals, which led to the design of tracking modules (illustrated in the emotion recognition system, FACS) and its relation to affect economies. Ultimately, I noticed that the better

the wider entanglement of agencies is understood – be it of the tracking module or of computational systems – the unlikelier it becomes to pinpoint any intentional moment or an agent position within these.

On Social Robots, Tracking, and Data Concerns as Ethically Entangled

Beginning this thesis, I assumed that data-related concerns around social robots were underestimated in Robot Ethics. However, I am assertive now that this neglect is deeply rooted in how the agency of robots is constituted. HR discusses tracking and tracking modules in social robots in a functional and neutral context of the HRI (Human-Robot-Interaction) only. Hence, I expanded this view ethically and aligned this process and interaction to Posthumanism and to Agamben's view on ethical technology, to allow for an alternative view on technology to emerge. This amplified the deeper reflection on agencies in tracking, but also left the focus on practical accountability open, even if tracking is better understood through a non-hierarchical approach. What must be taken from this theme is that I see no way around including the social robot – being a tracking device – in a critical discussion on data as undertaken in MSS, which can be understood better ethically and not morally.

On the Crucial Link between Robot Ethics and Dataveillance

The lack of reflection on the growing ubiquity of robots as a digital technology into new and often sensitive environments, such as elderly care, formed a crucial finding of my thesis' investigation, supported by the framework of (ii) Media and Surveillance Studies (MSS) and insights from Posthumanism. I anticipated that the theoretical negligence

around data in Robot Ethics would establish new ethical dilemmas on the sensitivity of patient and health data gathered by social robots, since social robots are underestimated in their potential position as dataveillance.

This dilemma stemmed from the conflicts I raised; the hybrid agency position and the conflation of moral and ethical concerns. This potential must not be taken lightly, as various scandals around the misuse of user data from social media platforms have shown recently that privacy is not a matter of an intention of a programmer or the platform owner, but it is a question of Big Data, information structures, commercial interests, and citizen literacy. I urged for Robot Ethics to liberate itself from HR goals and epistemes to some degree, since these focus on insufficient anthropomorphic agency models and questions on patient harm or isolation, but not on how these become increasingly autonomous data collectors.

If Robot Ethics remains unwilling to discuss the uncontrollable element in collection data, which is possible, then Media and Surveillance Studies must pick this up. These data concerns create crucial practical dilemmas, which sooner or later will endanger the safety of patient data and privacy, since the value of health and demographic data is growing for commercial industries. I stressed that, even if the concerns on social robots becoming potential dataveillance might be a future-led scenario, the technical possibility will be given soon and the implementation of robots into professional care will increase. Consequently, this will create new forms of dataveillance (or Big Data) structures. Therefore, I encouraged Robot Ethics to rethink their passive disciplinary

position around dataveillance and dismiss other views as disciplinary differences. The goal for which I advocate is to establish more clarity on data ownership, privacy, and agent literacy, and on the potential commodification of data when robots are applied.

On the Posthumanist Ethical Views on Robots

Reflecting on robots ethically was supported by Posthumanism and Agamben's ethical view on early cinema as a dispositif, which allowed me to challenge the morality-led discussion from Robot Ethics as anthropomorphic. Despite the usefulness in understanding tracking and data ethically, I developed a new awareness on why Posthumanist views might not be included in Robot Ethics (even if such might equally relate to a disciplinary ignorance).

First, certain progressive, ethical discussions – as sophisticated as these are – lacked practical applicability. However, their concerns were valid and essential to understanding that, as we debate agency models of technology and human, we cannot see emerging technology such as robots as being an entity or instrument, but must embrace their entanglement, fluidity, and performativity ethically. This view seemed impractical for the practical concerns of robots, since, by offering fluid and performative agency models, it did not allow to distinguish between agent positions practically. If everything is entangled, who or what is the moral agent?

Nonetheless, the lack of a practical applicability of Posthumanist theories is not a reason to exclude these from Robot Ethics, since these allowed for a better understanding of technology in terms of agency and ethical relations. I went as far as stating that not fully

understanding data collection, as enabled by the Posthumanist context, not only compromises the elderly's privacy and ownership concerns, but makes Robot Ethics an unethical practice.

As I left the frameworks of Robot Ethics and MSS, I entered (iii) Machine and Computer Ethics and wider Philosophy of Technology discourses on robots and machines, which focused on questions of moral agency, autonomy, and responsibility of technology as practically applied in digital, computational systems. These canons aligned moral questions to emerging technology's ability to judge, decide, or be autonomous as a responsible agency. The insights I gained from these discussions presented new advanced angles, but also offered limited translations of moral agency, which I labelled as leading to (2) Reductionist Morality and a (3) Distributed Morality views, with newly surfacing concerns on the revival of a moral positivism and on the lack of accountability questions in the use of robots, which I present next.

On Algorithmic and Computational Modelling of Moral Agency

The computational models of morality that I surveyed proved to be reductionist, context dependent, distributed, and/or operational. The concern I labelled as (2) Reductionist Morality is one on the aligning moral norms to a numeric, static, or universal set of rules. This process revealed various issues to me; for instance, a lacking critical reflection on how to reposition old concerns from moral philosophy into new POT contexts. As Robot Ethics and POT tended to choose specific thoughts and backdrops from traditional philosophical streams, such as Kantian, Utilitarian, or Humean ethics,

these incorporated the initial concerns these streams had not yet resolved (for instance, to differences between individual and collective ethics for a society, or on the sources or reasons for duty or moral virtues, etc.).

Therefore, what seemed insufficiently acknowledged is that traditional moral canons have tackled reductionist views on morality for centuries and, until now, there is no general agreement on what morally universal virtues or rules are that human agents should follow. Further, in recent Critical Theory and Posthumanism, multiple new and progressive view points and debates have emerged against any agency reductionism and hierarchy, presenting clearly how backwards and unsustainable stable, singular or rational models of human agency are.

However, POT streams are not always consistent or aligned in their reductionist morality and agency views; some have embedded certain affective or performative tendencies from Posthumanism to look at agency, but others are still very much dedicated to reinstating traditional and individual philosophy models of *the* agent, or *the* universal moral value (Floridi would align with the first, Wallach and Allen, with the second). By indirectly reviving Utilitarianism or Kantian ethics – both of which are concepts that allow for an action-led and rationally oriented moralism – reductionist POT views then support a mechanised view on human values, as much as on separated and compartmentalised agent entities.

From this discussion, I resumed that algorithmic morality has created another form of moral positivism, reviving a new institutionalisation of morality that is reminiscent of

how religious or legal morality is framed. Furthermore, this development does not give any opportunity for a new dynamic ethics to emerge, unfortunately. Reducing morality to universal values appeared not only to be dogmatic and to be reproducing or simplifying old moral concerns, but also lacked agreement on the reference points of *goodness* or *badness* in the moral evaluation. By following morality-led agent models, both Robot Ethics and POT laid their own traps. What began as their attempt to develop a *good* robot (be it superficially or computationally) exposed an inherent oversight on how this very attempt is always morally corrupted by whoever decided on its norms. Ultimately, morality appeared to be a fascinating theoretical vessel since, on the one hand, its fluidity and hollowness allowed for an ambiguity in undefined reference points, but, on the other, it always suggested an inherently value-oriented and human-made approval or judgement.

<u>On Discussing Moral Agency of Robots as a Context Issue</u>

The surfacing of morality concerns was, in many ways, limiting, but it is also only contextual. I noticed the tendency to debate moral concerns only in contexts with a likelihood for the robot to be exposed to an immoral or violent context. It emerged that robot morality research in POT concentrates on inherently morally conflicting contexts, such as military operations or drone applications, in which robots must ideally be able to judge and negotiate the level of creating or preventing harm. This is one reason why the social robot, when applied in elderly care, does not find much acknowledgment in

the (iii) Machine and Computer Ethics framework, but would find it if seen through (ii) MSS.

On Distributed Morality Models and Ethical Algorithms

Understanding computational and algorithmic structures in robots – less as individual or pseudo-agents, but as aligned with Posthumanist views as non-evaluative – enabled an enhanced and more complex ethical perspective on robots as digital technology. However, what I saw disappearing from this focus is the importance of the accountable agent in thinking of morality as a discourse of responsibility, consequence, and accountability. The second issue did not affect the theoretical value of the ethical complexity, but it proved insufficient for clarity on responsibility when using robots in elderly care, or in the use of digital technology.

Hence, viewing morality algorithmically as distributed was not seen as theoretically imprecise or flawed, as such. However, I debated it as practically insufficient. My problem stemmed from how human moral responsibility is transformed into an operational responsibility of systems, and the latter systematically replaced questions on practical accountability. One of my major issues towards an algorithmic morality understanding became, therefore, the detachment of morality from moral accountability, resulting from the dissolving of agenthood into agencies. This evolved because as these discussions assign more autonomy to digital systems, each of the theoretical canons agreed (or did not deny) that moral accountability is exclusively located in human agents. This agreement that technology can never be accountable raised the questions on

the purpose of debating machine morality in general. Floridi, for instance, detaches moral accountability from moral responsibility in order to operationalise algorithmic responsibility, yet frees it from any morally accountable connotation.

In socio-cultural human contexts, morality is not simply an action-based good/bad system; it has a purpose, because good and bad norms are installed for a reason so that agents will act according to them. The motivation to do so can stem from an individual fulfilment to be good (virtue) or from wanting to be good within a wider community. However, being morally good can also be understood as a process that only works because of facing the consequences of being punished and blamed for immorality. This moral responsibility, to bear consequences and to account for one's behaviour, is only understood by agents (be it of individual or legal bodies). Therefore, what Floridi suggested by operationalising moral responsibility in this 'mindless' morality model would leave the blame and punishment question unresolved within algorithms. I demand for a critical reconsideration of the purpose of morality if accountability plays no role in its constitution.

In my view, we (researchers, roboticists, philosophers, programmers, and industries) need to pay more attention on the implementation and conceptualisation of the term morality and its implied value as a regulatory system in society. It appeared to me as if most POT discussions avoided addressing accountability concerns critically, because accountability remains linked to practical responsibility and this aspect is very uncomfortable, since it raises various legal and policy conflicts around robots or

tracking systems. What these canons do, instead, is to focus on a segmentation of morality into attributes, which then are discussed metaphorically.

I found that discussions on moral appearance, moral agents, moral decisions, or moral rules in robots or algorithmic systems remain metaphorical to most extend but are not framed as such. Their metaphorical and abstracted level makes these discussions on the one hand, into complex and meta-ethical conversations, but on the other, to non-applicable equations for practical concerns, which is the opposite of what applied ethics around robotic and algorithmic agency should aim for.

On Dissolving of Agenthood, Accountability; and on the Purpose of Robot Morality

My investigation showed that algorithmic morality theories lacked any alignment to reward, punish, and blame discussions, especially after having abandoned views on an individual agenthood of digital systems. This way of thinking about technology always implies that the human agent is accountable, which is sensible in practical terms, but, considering the growth of operational autonomy in digital systems, complicates the human status enormously.

The unresolved question is still: How does a human agent deal with this new and complex collaboration with digital technology? In the context of robotics being a network of individuals programming, developing, and deciding, we need discussions on practical accountability as soon as possible. Also, what we consider moral behaviour is not a self-sufficient rule system outside of a real-life context of application, but it affects real human beings, such as the elderly, which are a specifically vulnerable group.

It proved challenging to fully resolve the conflicts I raised in the introduction, but I see the value of this thesis in having identified and debated them. It was not possible to simply fuse the mentioned moral and ethical streams together or to align their agency models theoretically, since their epistemological frameworks aim for different goals. However, fusing these streams into one investigative context justified why social robots require a new transdisciplinary ethics to resolve present disjoints, and this thesis offered the first step towards this becoming possible.

What I hopefully worked out throughout this thesis is that the use of digital technology has entangled complex questions of agenthood and superficial or ontological agency, of operational or moral intentions and actions, which have blurred any clear purpose of moral agency being necessary or possible in technologies such as robots. It has become difficult to pinpoint the value of a concept of morality and what the contemporary streams intend to achieve when conceptualising morality into digital systems. What my investigation illustrated was that the newly emerging socio-technological agencies in digital technology should not only be aligned to anthropocentric ways of searching for rational or locatable intentions, agents, or values in the technological system, but, as the segmentation of these discussion shows, the very use of robots must be challenged holistically.

Ultimately, what I urge for is to shift away from only focusing on how to make robots *more* moral or ethical and to examine the developer and programmers' ambitions in applying or designing robots. I did not imply to look for a stable and fixed human entity

or a rational subject who owns stable moral values, but I indicated that the human agent (as individual researcher or as collective industry) is (or must be) the decisive authority in the wider context of using robots. This is simply because robots do not actively care about their moral status or their practical consequences on the lives of people or society, but human agents are affected by this technology exclusively.

A common denominator of research should be to aim for progress that supports human wellbeing, instead of pushing for a greater and faster advancement of technology, even if this is where the funding and *innovation* is seen to be. Robots can be very useful technologies nonetheless, as cases show in which robots are assisting people who are physically impaired. However, the robotisation of care professions evolves quicker than the questions around its disadvantages can be answered.

Beyond the technological and commercial interests, placing more robots into professional care can, in my view, be used to mask another problematic tendency – the devaluation and exploitation of human caretakers. While the commercial interests are not as much scrutinised in ethical discussions as they should be, accusing roboticists, robot ethicists, or programmers of not having good intentions or not caring about what robots do would be unjust and one-dimensional. Likewise, I cannot expect that an individual researcher can change or account for a whole research stream. I do not speak from a technophobe point of view, but with a critical angle on hidden technocratic ambitions infiltrating (and the commodification of research on) human values. Consequentially, I also disapprove of new tendencies to debate robot rights before we

have agreed on how to protect human privacy rights and on how to sufficiently educate vulnerable and maybe less literate human groups, as they get exposed to robots and digital systems more frequently.

As a transdisciplinary researcher, I have shown that the ethical dimensions and consequences of using robots are debated, but often from various and unconnected angles. However, social robots (as much as digital technologies widely) are already and increasingly applied and tested in new environments aligned to care environments without an awareness of wider concerns. Researchers, like myself, must be critical and persistent in demanding more and clearer answers from HR and the industry, on who and what is profiting from the use of robots, to avoid the application of robots to become unethical in itself, and to avoid for morality to become an instrumentalised label.

The academic response to these challenges must not advocate *for* or *against* the developing of new technologies, such as robots, but must critique the commodification, the growing techno-dogmatism, and the renewal of moral positivism in this context. I am convinced that we have a choice to either allow for new research questions to be driven mainly by the industry and pushed through by technocratic visions of progress, or to decide on research goals driven by citizens' needs and values, which might require robots to assist, but hopefully not to care for us.

Further Research

I urge for further transdisciplinary research regarding data infringement and privacy issues around social robots and robotic technology in elderly care, on the connection between Robot Ethics to Data Ethics and on the limitations of integrating traditional moral philosophy into the research on Big Data (Herschel & Miori, 2017).

Further research into elderly care and social robots must address privacy and data issues rather than circle around anthropocentric privacy intrusion issues, as undertaken momentarily in Robot Ethics. The similarities between tracking modules in robots and website cookies was not addressed in this thesis, but influenced this thesis' trajectory, given it created my awareness around tracking and data infringement. Future research has to focus on an in-depth discussion of wider literature on website architectures and tracking applications such as 'cookies'[111] (Acar et al., 2014; Raley, 2012) to expand the critical knowledge on tracking applications. I was not able to fully explore this angle when reflecting on dataveillance in MSS. I highlight Sylvia's (2016) view on privacy protection as a *red herring* debate that masks 'a much larger argument about the changing character of the risks stemming from the power differential created by corporate control of information' (20).

---

[111] *Cookies* can be understood as tracking algorithms that create an online *fingerprint* of the behaviour of the web user. Such are increasingly difficult to avoid or disable due to newer versions such as *evercookies,* which link multiple levels of data access and pattern making (Acar et al., 2014). More on cookies and the tracking of cookie applications at: https://www.theguardian.com/technology/2012/apr/23/cookies-and-web-tracking-intro, https://www.npr.org/templates/story/story.php?storyId=129298003 (Accessed 26.02.2018).

Since robotics is an industry-driven research field (Rommetveit et al., 2012) with commercial interests, I stress again that every possible angle on data must be explored to understand its value as a commercialised commodity. Again, this problematic connection must consider wider ethics research as much as the shareholders and developers, in my view. For now, I showed that the concerns related to data are discussed in a limited way or are barely unpacked in Robot Ethics (exceptions in POT are Stahl & Coeckelbergh, 2016, and Knoppers & Thorogood, 2017). I advocated the association of Robot Ethics to Media and Surveillance Studies as a valued path, which enabled an association chain between social robots and potential data infringement through tracking, which can be picked up in the future.

I further advise for future research to explore the linkage between Robot Ethics and Data Ethics in much more detail than this thesis could, and to consider this link as an essential theory and practice in health contexts. Through the association of data-related issues, tracking, and social robots, I indirectly pointed to the connection between Robot Ethics and data-related ethical issues around the protection of health data, which becomes an increasingly commercialised topic (Knoppers & Thorogood, 2017). Floridi & Taddeo (2016) are critiquing the disconnection between Robot Ethics and Data Ethics, but their work is neither interested in what role tracking modules play, nor in social robots. I could not supply an in-depth discussion on Data Ethics, but Floridi's work was discussed in its scope of distributed morality.

I urge for further research on the linkage between machine morality and Big Data (synonymously used with dataveillance in this thesis) in newer POT streams, which try to move beyond traditional moral philosophy. The critique raised by Herschel & Miori (2017), for instance, shows why the alignment of traditional moral philosophies with Big Data structures is problematic. What I observed was that morality, as a concept, still dominates techno-philosophical discussions on machines and computers as autonomous technologies, while traditional moral philosophy is clearly considered to have a limited ability to grasp issues such as Big Data.

However, I noticed as much that Big Data research from POT often bypasses conversation about robots or angles from MSS. Floridi's work (2014) already undertook a first step in the right direction, in bringing down the view on computers as single agents, by offering the distributed model on moral agency. However, Floridi & Taddeo (2016) do not suggest or intend to overcome discussions on morality, as I pointed out, neither are they interested in social robots or elderly care – even if they do address the ethical consequences around robots and data by looking at different 'levels of abstraction' (LoA).[112] Hence, what I advocate as necessary is the exploration of robots and data/dataveillance/Big Data *through* the analysis of social robots and tracking, via Media and Surveillance Studies (MSS), to address newer ethical concerns.

[112] Floridi & Taddeo (2016) are not interested in the social robot and design implications. They only refer to robots as information structures.

Bibliography

Acar, G., Eubank, C., Englehardt, S. and Juarez, M. (2014). *The Web Never Forgets: Persistent Tracking Mechanisms in the Wild*. [online] kuleuven.be. Available at: https://securehomes.esat.kuleuven.be/~gacar/persistent/the_web_never_forgets.pdf (Accessed on 28.06.2018).

Admoni, H. & Scassellati, B. (2017). Social Eye Gaze in Human-Robot Interaction: A Review. *Journal of Human-Robot Interaction*, 6(1), p.25-63.

Adăscăliţei, F. and Doroftei, I. (2012). Expressing Emotions in Social Robotics - A Schematic Overview Concerning the Mechatronics Aspects and Design Concepts. *IFAC Proceedings Volumes*, 45(6), pp.823-828.

Agamben, G. (2000). *Means without end: Notes on Politics*. Minneapolis: University of Minnesota Press.

Agamben, G. (2009). *What is an Apparatus? And Other Essays*. Stanford: Stanford University Press.

Agamben, G. (2010). Notes on Gestures. In: G. Agamben, ed., *Means Without End: Notes on Politics*. Minneapolis: University of Minnesota Press, pp.49-63.

Ahmed, S. (2004). *The Cultural Politics of Emotion*. Edinburgh: Edinburgh University Press.

Andrejevic, M. & Burdon, M. (2014). Defining the Sensor Society. *Television & New Media*, 16(1), pp.19-36.

Andrejevic, M. (2017). *Toward a Program of Algorithmic Accountability: A Collaborative Approach*. [online] YouTube. Available at: https://www.youtube.com/watch?v=lkyLIayfdMU (Accessed on 25.05.2018).

Andrejevic, M. (2018). Ubiquitous Surveillance. In: K. Ball, K. Haggerty and D. Lyon, ed., *Routledge Handbook of Surveillance Studies*. Oxon: Routledge, pp.91-98.

Angerer, M., Bösel, B. & Ott, M. (2014). *Timing of Affect: Epistemologies, Aesthetics, Politics*. Chicago: University of Chicago Press.

Angerer, M. & Bösel, B. (2015). CAPTURE ALL, ODER: WHO'S AFRAID OF A PLEASING LITTLE SISTER?. *Zeitschrift für Medienwissenschaft ZFM*, 2(13), pp.48-56.

Arkin, R. (2009). Ethical Robots in Warfare. *IEEE Technology and Society Magazine*, 28(1), pp.30-33.

Ball, K., Haggerty, K. & Lyon, D. (2012). *Routledge Handbook of Surveillance Studies*. Oxon: Routledge.

Barad, K. (2003). Posthumanist Performativity: Toward an Understanding of How Matter Comes to Matter. *Signs: Journal of Women in Culture and Society*, 28(3), pp.801-831.

Barad, K. (2007). *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Durham: Duke University Press.

Bergo, B. (2011). *Emmanuel Levinas*. [online] Available at: https://plato.stanford.edu/entries/levinas/#OveLev (Accessed 10.01.2019).

Billard, A. & Dautenhahn, K. (1999). Experiments in Learning by Imitation - Grounding and Use of Communication in Robotic Agents. *Adaptive Behavior*, 7(3-4), pp.415-438.

Blackburn, S. (2001). *Ethics: A Very Short Introduction.* Oxford: Oxford University Press.

Blackman, L. (2012). *Immaterial Bodies: Affect, Embodiment, Mediation*. Los Angeles: SAGE Publications.

Braidotti, R. (2006). *Transpositions. On Nomadic Ethics.* Cambridge: Polity.

Braidotti, R. &Vermeulen, T. (2014). *Borrowed Energy*. [online] Frieze.com. Available at: https://frieze.com/article/borrowed-energy (Accessed on 15.06.2018).

Bray, H. (2014). *You Are Here. From the Compass to GPS, the History and Future of how we find ourselves*. Philadelphia: Basic Books.

Breazeal, C. (2002). *Designing Sociable Robots*. Cambridge, Mass.: MIT Press.

Breazeal, C. (2003). Emotion and Sociable Humanoid Robots. *International Journal of Human-Computer Studies*, 59(1-2), pp.119-155.

Breazeal, C., Edsinger, A., Fitzpatrick, P. & Scassellati, B. (2000). *Social Constraints on Animate Vision*. [online] Scazlab.yale.edu. Available at: https://scazlab.yale.edu/sites/default/files/files/Humanoids2000-vision.pdf (Accessed on 14. 07. 2018).

Brèthes, L., Lerasle, F. & Danès, P. (2008). Data Fusion for Visual Tracking dedicated to Human-Robot Interaction. In: *2005 IEEE International Conference on Robotics and Automation*. Barcelona: IEEE International Conference on Robotics and Automation, pp.2075-2080.

Brèthes, L., Menezes, P., Lerasle, E. & Hayet, J. (2018). Face Tracking and Hand Gesture Recognition for Human-Robot Interaction. In: *2004 IEEE International Conference on Robotics & Automation*. New Orleans: IEEE International Conference on Robotics & Automation, pp.1901-1906.

Brèthes, L., Menezes, P., Lerasle, F. & Hayet, J. (2004). Face tracking and hand gesture recognition for human-robot interaction. *Proceedings of the 2004 IEEE International Conference on Robotic Automation*, pp.1901-1906.

Brey, P. (2014). From Moral Agents to Moral Factors: The Structural Ethics Approach. In: P. Kroes and P. Verbeek, ed., *The Moral Status of Technical Artefacts*. Dordrecht: Springer Science+Business Media, pp.125-143.

Broadbent, E. (2017). Interactions with Robots: The Truths We Reveal About Ourselves. *Annual Review of Psychology*, 68(1), pp.627-652.

Brooks, R. (1991). *Intelligence Without Reason*. [online] People.csail.mit.edu. Available at: http://people.csail.mit.edu/brooks/papers/AIM-1293.pdf (Accessed on 13.07.2018).

Brooks, R. (2002). Humanoid Robots. *Communications of the ACM*, 45(3).

Bruce, A., Knight, J., Listopad, S., Magerko, B. & Nourbakhsh, I. (2000). Robot Improv: Using Drama to Create BelieKable Agents. *Proceedings of the 2000 IEEE International Conference on Robotics & Automation*, pp.4002-4006.

Bunge, M. (1977). Towards a Technoethics. *Monist,* 60 (1), pp. 96 - 107.

Cañamero, L. & Lewis, M. (2016). Making New "New AI" Friends: Designing a Social Robot for Diabetic Children from an Embodied AI Perspective. *International Journal of Social Robotics*, 8(4), pp.523-537.

Christians, C. (2011). Primordial Issues in Communication Ethics. In: R. Fortune and P. Fackler, ed., *The Handbook of Global Communication and Media Ethics*, 1st ed. Chichester: Blackwell Publishing, pp.1-20.

Clarke, R. (1988). Information Technology and Dataveillance. *Communications of the ACM*, 31(5), pp.498-512.

Clemens, J., Heron, N. & Murray, A. (2011). *The Work of Giorgio Agamben*. Edinburgh: Edinburgh University Press.

Coeckelbergh, M. (2009). Virtual Moral Agency, Virtual Moral Responsibility: On the Moral Significance of the Appearance, Perception, and Performance of Artificial Agents. *AI & SOCIETY*, 24(2), pp.181-189.

Coeckelbergh, M. (2010). Moral Appearances: Emotions, Robots, and Human Morality. *Ethics and Information Technology*, 12(3), pp.235-241.

Couldry, N. (2012). *Media, Society, World: Social Theory and Digital Media Practice*. Cambridge: Polity.

Couldry, N., Madianou, M. & Pinchevski, A. (2013). *Ethics of Media*. Basingstoke: Palgrave Macmillan.

Damiano, L. & Dumouchel, P. (2018). Anthropomorphism in Human–Robot Co-evolution. *Frontiers in Psychology*, 9(468), pp.1-9.

Dario, P., Guglielmelli, E., Laschi, C. & Teti, G. (1999). MOVAID: A Personal Robot in Everyday Life of Disabled and Elderly People. *Technol. Disabil.*, 10, pp.77–93.

Davidson, R., Sherer, K. & Goldsmith, H. (2003). *Handbook of Affective Sciences*. New York, NY: Oxford University Press.

Davisson, A. & Booth, P. (2016). *Controversies in Digital Ethics*. New York: Bloomsbury Academic.

De Greeff, J. & Belpaeme, T. (2015). Why Robots Should Be Social: Enhancing Machine Learning through Social Human-Robot Interaction. *PLOS ONE*, 10(9), pp. 1-26.

Didi-Huberman, G. (2004). *Invention of Hysteria: Charcot and the Photographic Iconography of the Salpetriere*. Cambridge, MA: The MIT Press.

Dimitrova, Z. (2017). Robotic Performance: An Ecology of Response. *Performance Philosophy*, 3(1), pp.162-177.

Dodig Crnkovic, G. & Çürüklü, B. (2012). Robots: Ethical by Design. *Ethics and Information Technology*, 14(1), pp.61-71.

Draper, H. & Sorell, T. (2017). Ethical Values and Social Care Robots for Older People: An International Qualitative Study. *Ethics and Information Technology*, 19(1), pp.49-68.

Drushel, B. & German, K. (2011). *The Ethics of Emerging Media: Information, Social Norms, and New Media Technology*. New York: Continuum.

Duffy, B. (2003). Anthropomorphism and the Social Robot. *Robotics and Autonomous Systems*, 42(3-4), pp.177-190.

Dumouchel, P. & Damiano, L. (2017). *Living with Robots*. Cambridge: Harvard University Press.

Durantin, G., Heath, S. & Wiles, J. (2017). Social Moments: A Perspective on Interaction for Social Robotics. *Frontiers in Robotics and AI*, 4.

Ekman, P. (2003). *Emotions Revealed: Recognising Faces and Feelings to Improve Communication and Emotional Life*. New York: Times Books.

Elmer, G. (2012). Panopticon—Discipline—Control. In: K. Ball, K. Haggerty and D. Lyon, ed., *Routledge Handbook of Surveillance*. Oxon: Routledge, pp.21-30.

Engelberger, J. (1989). *Robotics in Service*. London: Kogan Page.

Eshleman, A. (2014). *Moral Responsibility*. [online] Plato.stanford.edu. Available at: https://plato.stanford.edu/entries/moral-responsibility/ (Accessed on 19.07.2018).

Floridi, L. (2010). *The Cambridge Handbook of Information and Computer Ethics*. Cambridge: Cambridge University Press.

Floridi, L. (2014). Artificial Agents and Their Moral Nature. In: P. Kroes and P. Verbeek, ed., *The Moral Status of Technical Artefacts*. Dordrecht: Springer Science+Business Media, pp.185-213.

Floridi, L. & Taddeo, M. (2014). *The Ethics of Information Warfare*. Cham: Springer International Publishing.

Floridi, L. & Taddeo, M. (2016). What is Data Ethics?. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083), p.20160360.

Flusser, V. (1993). *Lob der Oberflächlichkeit: für eine Phänomenologie der Medien*. Mannheim: Bollmann.

Flusser, V. (2000). *Towards a Philosophy of Photography*. London: Reaktion Books.

Flusser, V. & Roth, N. (2014). *Gestures*. Minneapolis: University of Minnesota Press.

Fong, T., Nourbakhsh, I. & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42(3-4), pp.143-166.

Ford, C. (2014). *Blinking in Human Communicative Behaviour and its Reproduction in Artificial Agents*. Doctor of Philosophy. Plymouth University.

Fortner, R. & Fackler, P. ed., (2011). *The Handbook of Global Communication and Media Ethics*. 1st ed. Chichester: Blackwell Publishing.

Galloway, A. (2012). *The Interface Effect*. Cambridge, UK: Polity Press.

Gates, K. (2011). *Our Biometric Future: Facial Recognition Technology and the Culture of Surveillance*. New York: NEW YORK UNIVERSITY PRESS.

Gawne, M. (2012). The Modulation and Ordering of Affect: From Emotion Recognition Technology to the Critique of Class Composition. *Exploring Affective Interactions*, (21), pp.98-124.

Gitelman, L. (2013). *"Raw Data" Is an Oxymoron*. Cambridge: MIT Press.

Gregg, M. & Seigworth, G. (2010). *The Affect Theory Reader*. Durham: Duke University Press.

Guizzo, E. (2015). *Jibo Is as Good as Social Robots Get. But Is That Good Enough?* [online] IEEE Spectrum: Technology, Engineering, and Science News. Available at: https://spectrum.ieee.org/robotics/home-robots/jibo-is-as-good-as-social-robots-get-but-is-that-good-enough (Accessed on 20.04.2018).

Gunkel, D. (2016). Paradigm Shift: Media Ethics in the Age of Intelligent Machines. In: A. Davisson and P. Booth, ed., *Controversies in Digital Ethics*. New York: BLOOMSBURY, pp.233-248.

Gustafsson, H. & Grønstad, A. (2014). *Cinema and Agamben: Ethics, Biopolitics and the Moving Image*. New York: Bloomsbury Academic.

Hackel, M. (2007). *Humanoid Robots: Human-like Machines*. Vienna: I-Tech Education and Publishing.

Hall, L. (2017). *How We Feel About Robots That Feel*. [online] MIT Technology Review. Available at: https://www.technologyreview.com/s/609074/how-we-feel-about-robots-that-feel/ (Accessed on 11.04.2018).

Harbord, J. (2016). *Ex-centric cinema: Giorgio Agamben and Film Archaeology*. New York: Bloomsbury Academic.

Haritaoglu, I., Harwood, D. and Davis, L. (2000). W4: Real-Time Surveillance of People and their Activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), pp.809-830.

Hay, M. (2014). *Could a Machine feel Human-like Emotions?* [online] Life 2.0. Available at: https://www.vitamodularis.org/articles/could_a_machine_feel_human-like_emotions.shtml (Accessed on 16.05.2018).

Hepp, A. & Krotz, F. (2014). *Mediatized Worlds: Culture and Society in a Media Age*. Basingstoke: Palgrave Macmillan.

Herschel, R. & Miori, V. (2017). Ethics & Big Data. *Technology in Society*, 49, pp.31-36.

Höök, K. (2009). Affective Loop Experiences: Designing for Interactional Embodiment. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535), pp.3585-3595.

Horvath, S. (2013). *Aktueller Begriff: Big Data*. [online] Bundestag. Available at: https://www.bundestag.de/blob/194790/c44371b1c740987a7f6fa74c06f518c8/big_data-data.pdf (Accessed on 15.02.2018).

Introna, L. & Wood, D. (2004). Picturing Algorithmic Surveillance: The Politics of Facial Recognition Systems. *Surveillance & Society*, 2(2/3), pp.177-198.

Jain, B. (2012). Efficient Iris Recognition Algorithm Using Method of Moments. *International Journal of Artificial Intelligence & Applications*, 3(5), pp.93-105.

Johansson, L. (2011). *Robots and Moral Agency.* Licentiate Thesis. Royal Institute of Technology Stockholm.

Johnson, D. & Noorman, M. (2014). Artefactual Agency and Artefactual Moral Agency. In: P. Kroes and P. Verbeek, ed., *The Moral Status of Technical Artefacts*. Dordrecht: Springer Science+Business Media, pp.143-159.

Jokinen, K. (2009). Gaze and Gesture Activity in Communication. In: *Universal Access in Human-Computer Interaction*. San Diego: Proceedings of the 5th International Conference, pp.1-11.

Kammerer, D. & Waitz, T. (2015). ÜBERWACHUNG UND KONTROLLE. *Zeitschrift für Medienwissenschaften*, (13).

Kahn, Jr, P., Kanda, T., Ishiguro, H. and Gill, B. (2012). Do People Hold a Humanoid Robot Morally Accountable for the Harm It Causes?. *HRI'12*, pp.33-40.

Kanda, T. & Ishiguro, H. (2013). *Human-Robot Interaction in Social Robotics*. Boca Raton: CRC Press.

Kanda, T., Iwase, K., Shiomi, M. & Ishiguro, H. (2013). Moderating Users' Tension to Enable Them to Exhibit Other Emotions. In: T. Kanda and H. Ishiguro, ed., *HUMAN–ROBOT INTERACTION IN SOCIAL ROBOTICS*. Boca Raton: CRC Press, pp.299-311.

Kendon, A. (2004). *Gesture. Visible Action as Utterance*. New York: Cambridge University Press.

Knapp, M., Hall, J. & Horgan, T. (2014). *Nonverbal Communication in Human Interaction*. 8th ed. Boston: Wadsworth.

Kirby, R., Forlizzi, J. & Simmons, R. (2010). Affective Social Robots. *Robotics and Autonomous Systems*, 58(3), pp.322-332.

Knoppers, B. & Thorogood, A. (2017). Ethics and Big Data in health. *Current Opinion in Systems Biology*, 4, pp.53-57.

Kroener, I. & Neyland, D. (2012). New technologies, security and surveillance. In: K. Ball, K. Haggerty and D. Lyon, ed., *Routledge Handbook of Surveillance Studies*. Oxon: Routledge, pp.141-149.

Kroes, P. &Verbeek, P. ed., (2014). *The Moral Status of Technical Artefacts*. Dordrecht: Springer Science+Business Media.

Krotz, F. (2012). Intimate Communication on the Internet: How Digital Media are Changing our Lives at the Microlevel. In: E. Wyss, ed., *Communication of Love: Mediatized Intimacy from Love Letters to SMS. Interdisciplinary and Historical Studies*. Bielefeld: transcript Verlag, pp.79-92.

Laidlaw, J. (2014). The Undefined Work of Freedom: Foucault's Genealogy and the Anthropology of Ethics. In: J. Faubion, ed., *Foucault Now: Current Perspectives in Foucault Studies*. Cambridge: Polity, pp.23-37.

Laney, D. (2001). *3D Data Management: Controlling Data Volume, Velocity, and Variety*. [online] Blogs.gartner.com. Available at: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (Accessed on 20.04.2018).

Lee, K., Park, N. & Song, H. (2005). Can a Robot Be Perceived as a Developing Creature?. *Human Communication Research*, 31(4), pp.538-563.

Leite, I., Pereira, A., Mascarenhas, S., Martinho, C., Prada, R. & Paiva, A. (2013). The Influence of Empathy in Human–Robot Relations. *International Journal of Human-Computer Studies*, 71(3), pp.250-260.

Lemaignan, S., Fink, J., Dillenbourg, P. & Braboszcz, C. (2014). The Cognitive Correlates of Anthropomorphism. *Proceedings Human-Robot Interaction Conference*. [online] Available at: https://infoscience.epfl.ch/record/196441/files/anthropomorphism-cognition.pdf; (Accessed on 10.01.2018).

Levinas, E. & Hand, S. (1989). *The Levinas Reader: Emmanuel Levinas*. Oxford: Blackwell.

Levitt, D. (2011). Notes on Media and Biopolitics: 'Notes on Gesture'. In: J. Clemens, N. Heron and A. Murray, ed., *Work of Giorgio Agamben: Law, Literature, Life: Law, Literature, Life*. Edinburgh: Edinburgh University Press, pp.193-212.

Levy, D. (2009). *Love and sex with robots: the evolution of human-robot relationships*. London: Duckworth Overlook.

Lin, P., Bekey, G. and Abney, K. (2008). *Autonomous Military Robotics: Risk, Ethics, and Design*. Ft. Belvoir: Defense Technical Information Center.

Lin, P., Abney, K. & Bekey, G. (2011). Robot Ethics: Mapping the Issues for a Mechanized World. *Artificial Intelligence*, 175(5-6), pp.942-949.

Lin, P., Abney, K. & Bekey, G. (2012). *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge: MIT Press.

Lipton, Z. (2017). *Press Failure: The Guardian's "Meet Erica"*. [online] Approximatelycorrect.com. Available at: http://approximatelycorrect.com/2017/04/17/press-failure-guardian-meet-erica/ (Accessed on 10.02.2018).

Luhmann, N. (1986). *Love as Passion: The Codification of Intimacy*. Cambridge: Harvard University Press.

Luppicini, R. (2009). *The emerging field of Technoethics* [online] Available at: https://bit.ly/2RW7CaR (Accessed on 12.01.2019)

Luhmann, N. (1989). *Ecological Communication*. Chicago: University of Chicago Press.

Lundby, K. (2014). Mediatized Stories in Mediatized Worlds. In: A. Hepp and F. Krotz, ed., *Mediatized Worlds: Culture and Society in a Media Age*. Palgrave Macmillan, pp.19-38.

Malle, B. & Scheutz, M. (2015). When Will People Regard Robots as Morally Competent Partners?. In: *24th IEEE International Symposium on Robot and Human Interactive Communication*. Kobe: IEEE International, pp.486-491.

Manovich, L. (1996). *ON TOTALITARIAN INTERACTIVITY (Notes from the Enemy of the People)*. [online] Manovich. Available at: http://manovich.net/content/04-projects/017-on-totalitarian-interactivity/14_article_1996.pdf (Accessed on 21.07. 2018).

Massumi, B. (2002). *Parables for the Virtual: Movement, Affect, Sensation*. Durham: Duke University Press.

Menezes, P., Lerasle, F., Dias, J. & Germa, T. (2007). Towards an Interactive Humanoid Companon with 367 Visual Tracking Modalities. In: M. Hackel, ed., *Humanoid Robots Human-like Machines*. Vienna: I-Tech Education and Publishing, pp.367-399.

Miller, S. (2008). Collective Responsibility and Information and Communication Technology. In: Van den Hoven, J. & Weckert, J. (2008). *Information Technology and Moral Philosophy*. Cambridge University Press, pp. 226- 251.

Moeller, H. (2006). *Luhmann Explained: From Souls to Systems*. Chicago: Open Court Publishing Company.

Mori, M., MacDorman, K. & Kageki, N. (2012). The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine*, 19(2), pp.98-100.

Mosier, K., Skitka, L., Heers, S. & Burdick, M. (1998). Automation Bias: Decision Making and Performance in High-Tech Cockpits. *The International Journal of Aviation Psychology*, 8(1), pp.47-63.

Moubayed, S., Edlund, J. & Beskow, J. (2012). Taming Mona Lisa: Communicating Gaze faithfully in 2D and 3D Facial Projections. *ACM Transactions on Interactive Intelligent Systems*, 1(2), pp.1-25.

Münsterberg, H. & Langdale, A. (2002). *Hugo Münsterberg on Film: The Photoplay: A Psychological Study and Other Writings*. New York: Routledge.

Nilsson, N. (1984). *Shakey the Robot*. Menlo Park: SRI International. Artificial Intelligence Center.

Noorman, M. (2018). *Computing and Moral Responsibility*. [online] Plato.stanford.edu. Available at: https://plato.stanford.edu/entries/computing-responsibility/index.html (Accessed on 30.06.2018).

Norris, C. &Armstrong, G. (1999). *The Maximum Surveillance Society*. Oxford: Berg.

Northcott, P. (1998). *The Image of Hypnosis: Strange Beliefs, Strange Contexts, Familiar Behaviours*. Doctor of Philosophy. Bristol University.

Paiva, A., Leite, I. & Ribeiro, T. (2014). *Emotion Modelling for Social Robots*. [online] People.ict.usc.edu. Available at: http://people.ict.usc.edu/~gratch/CSCI534/Readings/ACII-Handbook-Robots.pdf (Accessed on 07.06.2017).

Paiva, A., Leite, I., Boukricha, H. & Wachsmuth, I. (2017). Empathy in Virtual Agents and Robots. *ACM Transactions on Interactive Intelligent Systems*, 7(3), pp.1-40.

Parikka, J. (2014). *Media Archaeology Out of Nature: An Interview with Jussi Parikka. E-flux*. [online] E-flux.com. Available at: https://www.e-flux.com/journal/62/60965/media-archaeology-out-of-nature-an-interview-with-jussi-parikka/ (Accessed on 30.06.2018).

Parisi, D. & Petrosino, G. (2010). Robots that have Emotions. *Adaptive Behavior*, 18(6), pp.453-469.

Parisi, L. (2013). *Contagious Architecture*. Cambridge, MA: The MIT Press.

Pelachaud, C. and Poggi, I. (2002). Multimodal Embodied Agents. *The Knowledge Engineering Review*, 17(02), pp.181–196.

Pias, C. (2018). *On the Epistemology of Computer Simulation*. [online] Genealogy-of-media-thinking. Available at: http://genealogy-of-media-thinking.net/wp-content/uploads/2013/06/CP0003.pdf (Accessed on 01.06.2018).

Pötzsch, H. (2017). *Media Matter*. [online] TripleC. Available at: https://www.triple-c.at/index.php/tripleC/article/view/819 (Accessed on 10.02.2018).

Prakash, J., Swami, P., Khandelwal, G., Singh, M. & Vijayvargiya, A. (2016). Digitally Transparent Interface Using Eye Tracking. *Procedia Computer Science*, 84, pp.57-64.

Punt, M. (2000). *Early Cinema and the Technological Imaginary*. Trowbridge: Cromwell Press.

Püschel, F. (2014). *Big Data und die Rückkehr des Positivismus. Zum gesellschaftlichen Umgang mit Daten*. [online] Medialekontrolle. Available at: http://www.medialekontrolle.de/wp-content/uploads/2014/09/Pueschel-Florian-2014-03-01.pdf (Accessed on 10.04. 2018).

Quan, W., Niwa, H., Ishikawa, N., Kobayashi, Y. & Kuno, Y. (2011). Assisted-care Robot based on Sociological Interaction Analysis. *Computers in Human Behavior*, 27(5), pp.1527-1534.

Raley, R. (2013). Dataveillance and Countervailance. In: L. Gitelman, ed., *"Raw Data" Is an Oxymoron*. Cambridge: MIT Press, pp.121-147.

Ravetto-Biagioli, K. (2016). The Digital Uncanny and Ghost Effects. *Screen*, 57(1), pp.1-20.

Read, R. (2014). *A Study of Non-Linguistic Utterances for Social Human-Robot Interaction*. Doctor of Philosophy. University of Plymouth.

Reeves, B. and Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge: Cambridge University Press.

Reilly, K. (2011). *Automata and Mimesis on the Stage of Theatre History*. Basingstoke: Palgrave Macmillan.

Robin, R. (2014). *A Study of Non-Linguistic Utterances for Social Human-Robot Interaction*. Plymouth: Plymouth University.

Roger, K., Guse, L., Mordoch, E. & Osterreicher, A. (2012). Social Commitment Robots and Dementia. *Canadian Journal on Aging / La Revue canadienne du vieillissement*, 31(01), pp.87-94.

Rossini, N. (2012). *Reinterpreting Gesture as Language*. Amsterdam: IOS Press.

Royakkers, L. & van Est, R. (2016). *Just Ordinary Robots. Automation from Love to War.* Boca Raton: CRC Press.

Ruprecht, L. (2010). Ambivalent Agency: Gestural Performances of Hands in Weimar Dance and Film. *A Journal of Germanic Studies*, 46(3), pp.255-275.

Ruprecht, L. (2017). Introduction: Towards an Ethics of Gesture. *Performance Philosophy*, 3(1), p.4.

Schermer, B. (2007). *Software Agents, Surveillance, and the Right to Privacy: A Legislative Framework for Agent-enabled Surveillance (SIKS dissertation series, 1873-0760; no. 2007-05)*. Amsterdam: Amsterdam University Press.

Scheutz, M. (2012). The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots. In: P. Lin, K. Abney and G. Bekey, ed., *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge: MIT Press, pp.205-223.

Sharkey, A. & Sharkey, N. (2010). Granny and the Robots: Ethical Issues in Robot Care for the Elderly. *Ethics and Information Technology*, 14(1), pp.27-40.

Sharkey, A. (2014). Robots and Human Dignity: A Consideration of the Effects of Robot Care on the Dignity of Older People. *Ethics and Information Technology*, 16(1), pp.63-75.

Sharkey, N. (2009). The Robot Arm of the Law Grows Longer. *Computer*, 42(8), pp.116-115.

Sharkey, N. (2012). Killing Made Easy: From Joysticks to Politics. In: P. Lin, K. Abney and G. Bekey, ed., *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge: MIT Press, pp.111-129.

Shukla, M. & Shukla, A. N. (2012). Growth of Robotics Industry Early in 21st Century. *In- ternational Journal of Computational Engineering Research (IJCER), 2(5), pp*. 1554-1558.

Singh, A. (2015). *'Pepper' the emotional robot, sells out within a minute'. CNN*. [online] Edition CNN. Available at: https://edition.cnn.com/2015/06/22/tech/pepper-robot-sold-out/index.html (Accessed on 18.07.2018).

Sparrow, R. & Sparrow, L. (2006). In the Hands of Machines? The Future of Aged Care. *Minds & Machines*, 16 (2), pp.141-162.

Sparrow, R. (2015). Robots in Aged Care: A Dystopian Future?. *AI & SOCIETY*, 31(4), pp.445-454.

Spencer, M. (2012). *Reason and Representation in Scientific Simulation*. Doctor of Philosophy. Goldsmith University.

Stahl, B. & Coeckelbergh, M. (2016). Ethics of Healthcare Robotics: Towards Responsible Research and Innovation. *Robotics and Autonomous Systems*, 86, pp.152-161.

Stamboliev, E. (2017). On *Spillikin – A Love Story*: Issues around the Humanoid Robot as a Social Actor on Stage, *AVANT*, *8*, Special Issue, pp.265–271.

Stamboliev, E. & Jackson, A., (2018). Human(oid) Devices. Instrumentation of Physicality in Performance. *Transtechnology Research Reader 2015-2017*, pp. 90-99.

Sylvia, J. (2016). Little Brother: How Big Data Necessitates an Ethical Shift from Privacy to Power. In: A. Davisson and P. Booth, ed., *Controversies in Digital Ethics*. New York: Bloomsbury Academic, pp.13-29.

Telotte, J. (1995). *Replications: A Robotic History of the Science Fiction Film*. Urbana: University of Illinois Press.

Tilley, E. (2011). New Culture/Old Ethics: What Technological Determinism Can Teach Us About New Media And Public Relations Ethics 191. In: B. Drushel and K. German, ed., *The Ethics of Emerging Media: Information, Social Norms, and New Media Technology*. New York: Continuum International Publishing Group, pp.191-213.

Tomkins, S. (2008). *Affect Imagery Consciousness: The Complete Edition*. New York: Springer.

Tseng, S., Chao, Y., Lin, C. & Fu, L. (2016). Service Robots: System Design for Tracking People through Data Fusion and Initiating Interaction with the Human Group by Inferring Social Situations. *Robotics and Autonomous Systems*, 83, pp.188-202.

Turkle, S. (2005). Relational Artifacts/Children/Elders: The Complexities of Cybercompanions. *Proceedings of the CogSci Workshop on Android Science*, pp.62-73.

Turkle, S. (2007). Authenticity in the Age of Digital Companions. *Interaction Studies*, 8(3), pp.501-517.

Turkle, S. (2011). *Alone Together*. New York: Basic Books.

Väliaho, P. (2010). *Mapping the Moving Image: Gesture, Thought and Cinema circa 1900*. Amsterdam: Amsterdam University Press.

Van de Poel, I. & Royakkers, L. (2006). The Ethical Cycle. *Journal of Business Ethics*, 71(1), pp.1-13.

Van den Hoven, J. (2010). The use of Normative Theories in Computer Ethics. In: L. Floridi, ed., *The Cambridge Handbook of Information and Computer Ethics*. Cambridge: Cambridge University Press, pp.59-77.

Van den Hoven, J. & Weckert, J. (2008). *Information Technology and Moral Philosophy*. Cambridge: Cambridge University Press.

Wallach, W. (2010). Robot Minds and Human Ethics: The Need for a Comprehensive Model of Moral Decision Making. *Ethics and Information Technology*, 12(3), pp.243-250.

Wallach, W. & Allen, C. (2009). *Moral Machines. Teaching Robots Right from Wrong.* Oxford: Oxford University Press.

Ward, S. (2015). *Radical Media Ethics: A Global Approach*. Chichester: John Wiley & Sons.

Winner, L. (1977). *Autonomous Technology and Political Thought: Technics-out-of-Control as a Theme in Political Thought*. Cambridge: MIT Press.

Wu, Y., Fassert, C. & Rigaud, A. (2012). Designing Robots for the Elderly: Appearance Issue and beyond. *Archives of Gerontology and Geriatrics*, 54(1), pp.121-126.

Xie, M. (2003). *Fundamentals of Robotics: Linking Perception to Action*. New Jersey: World Scientific Pub.

Ziafati, P. (2016). *Social Robots – Programmable by Everyone*. [online] University of Luxembourg. Available at: https://wwwen.uni.lu/university/news/latest_news/social_robots_programmable_by_everyone (Accessed on 20.06.2018).

Zielinski, S. (2013). *[...After the media] News from the Slow-Fading Twentieth Century*. Minneapolis, MN: Univocal Publishing.

Zylinska, J. (2014). *Minimal Ethics for the Anthropocene*. Ann Arbor: Open Humanities Press.