MASTER THESIS

THE PORTRAYAL OF ARTIFICIAL INTELLIGENCE IN SCIENCE FICTION LITERATURE

AN INVESTIGATION OF STEREOTYPES, TROPES AND GENDER

presented by

Larissa Haas

Raiffeisenstr. 35
67489 Kirrweiler

Mannheim Master in Data Science
5th Semester
Student ID: 1417669
larhaas@mail.uni-mannheim.de

submitted to

Prof. Dr. Heiko Paulheim
Chair for Data Science
School of Business Informatics and Mathematics
University of Mannheim

second opinion by

Annika Gonnermann
Anglistik II

supported by

Dr. Regina Schober
Anglistik III
Anglistisches Seminar
University of Mannheim

July 19, 2019

Abstract

Artificial intelligences (AI) inhabit our daily lives, but we do not know how this living together will look like in the future.

Stories from Science Fiction (SF) show different futures and different theories about the relationships between humans and technology. With the analysis of an extensive list of 44 stories about AIs and the surrounding society, we can answer questions about the future and our understanding of AI today.

The results show that the stories present AIs more often as male than as female. However, female AIs are still more connected to the picture of serving AIs than male AIs are. Technical programs are pushed into the gender stereotypes that we want to abolish in our societies today. To change the discrepancy between the presentations of gender roles and other typical tropes in SF stories, we need to be aware of them. This study aims at raising this awareness for readers and developers.

Contents

List of Figures

List of Tables

Chapter 1

AIs - Our Daily Companions

Artificial intelligences (AI) inhabit our daily lives, for example in the form of virtual assistants like *Siri*, *Alexa*, *Cortana*, and Co. However, at the same time, non-virtual forms are becoming more popular. We find vacuum cleaning and grass cutting robots in many households already. Based on those simple forms of AI, researcher are aiming for more: more autonomy, more variability, more intelligence. Is this development going to change the way we live with this AIs? Will it change our view on technology, on our society, maybe even on humanity itself?

We cannot know where the development will lead us. We cannot sneak-peak into the future to know which questions and problems will be essential at some point in time. However, people are imagining the future and thinking about the cohabitation of AI and humanity, Science Fiction authors in particular. Based on the recent understanding of technology and society, Science Fiction (SF) tries to aspire our life in the future, and AIs are common elements in this imaginations.

To answer this questions, and maybe to have a sneak-peak at the future, how we may live together with non-human consciousness, this paper will first dive deep into the research of artificial intelligences, in reality as well as in Science Fiction literature. It will then formulate expectations about the AIs found in SF, about their appearance, their behavior, their roles, and the stereotypes and tropes connected to them. In the end, the paper will have a look at selected SF novels and novellas. The focus will be on the portrayal of AIs: How is the relationship between humans and technology described? Has the portrayal changed over the last years?

Chapter 2

The Technological View on AI

In the research field of AI, there has been more progress in the last five years than in the 50 years before (Eberl, 2018). New fields of application are emerging; the shapes and dimensions of AI are changing. Accordingly, the picture of AI in our awareness is changing as well. In the following chapter, we will throw a glance at the research field of AI. We will learn about what it is and how it evolved to get a broader view of their variances. We will then look at how humans interact with technology, at the expectations we have and at the implications of that interaction.

2.1 Current AI Research

First, we will have a look at AI research itself. The following sections will examine the different notions of AI as well as the history of research. Afterward, we will take a glimpse at the AIs in our everyday lives and at the possibility that they, or something completely different, will turn into a superintelligence.

Definition

In general, the term "artificial intelligence" is hard to grasp. At its first usage, by John McCarthy at a conference in 1960, it was used to describe "the effort to make computers think" (Moravec, 1988, p. 8). Though it sounded like Science Fiction, from now on the term began to evolve into something real. Today, AI is a field of research and applied technology, and a buzzword widely recognized and requested.

For the term AI, we find mainly two different notions of understanding. On the one hand, the term describes a sub-part of research and technology. Through building on significant findings from Big Data and statistics, *Machine Learning* algorithms extract specific patterns from large amounts of data. The Cambridge En-

glish Dictionary defines AI as "the use of computer programs that have some of the qualities of the human mind, such as the ability to understand language, recognize pictures, and learn from experience" (CED: Artificial Intelligence, 2019). However, even this dictionary definition leaves much space for interpretation. Some authors note that "[e]ven in the narrowest, engineering-centric definitions, AI refers to a broad constellation of computing technologies" (Fast and Horvitz, 2017, p. 1), and those technologies have application scopes in various contexts and situations.

On the other hand, taken in its literal sense, AI is a human-like intelligence, created not by natural evolution and growing up, but *artificially*. This intelligence can occur, for example, by constructing it from scratch, copying it from existing intelligences (such as humans) or evolving on its own, probably in a digital environment. We will learn more about creating an AI in the remaining parts of this chapter.

These two above mentioned approaches do not conflict with each other, they just focus on different aspects of AI: The first approach, it will be called the *technical approach*, takes more specific abilities into account, such as text understanding or situational learning, the second one, which will be called the *holistic approach* in the following paragraphs, focuses on a more complete picture of intelligence.

The problem is mainly in the term *intelligent*. When a program is trained to decide if the picture shows a dog or not, this performance may appear to us humans as intelligent, because the program does an excellent job compared to other living beings other than humans. But when we think about how the program comes to its decisions (by processing millions of dog and none-dog pictures, by calculating weights for specific pixel constellations and so on), and what this program is able to do (recognize dogs, not cats, cars, humans, or anything else except for dogs), the impression of "intelligence" fades. This program would pass in terms of the technical approach because it can "recognize pictures", whereas it may not be sufficient for the holistic approach.

Moravec (1999) also thought about this problem. He observes, the more we knew about programs and computing, the more the computer had to do to prove its intelligence. Software engineers would only see the code that produces the result and would not see anything *intelligent* within. They knew the program decides based on probabilities and calculations. Accordingly, the human brain would not reveal its intelligence under the microscope. There the observer can only see cells and matter, and nothing that shows intelligence (Moravec, 1999, pp. 110-111). Then, what is intelligent? An interpretation, a characteristic only granted by and for humans?

Milestones in the Past and Today

Before we talk about this question, we will first have a look at the fundamentals of the term AI. Up to this point in time, the history of research about AI lived through several ups and downs of success and associated popularity, where the downs were called *AI winter* to show that scientists and developers put all plans and experiments on ice.

The first AI winter was observable in the late 70s. During the years before, AI had gained more and more interest because of first successful experiments with mathematical and logical problem-solving. Those applied situations had been relatively small and clear-cut, and the algorithms had succeeded by testing all possible solutions and giving the best as a result. With the 1970s these situations grew bigger and bigger, the computer scientists noticed that storing capacity and processor speed were not sufficient for the explosively increasing complexity of the problems they wanted to solve (Bostrom, 2014, p. 20).

In the 1980s, a melting was noticeable: Developers, mainly from Japan, created *expert systems*. These rule-based algorithms were able to deduce solutions from data based on rules. They were used to facilitate decision finding, but they were quite expensive in the creation and error-prone because the data sets and rules had to be created by humans by hand (Bostrom, 2014, p. 21). This disenchantment led to the *second AI winter* which lasted until the 1990s. Now the processors and the storage capacity were large enough to process enough data for more sophisticated algorithms and approaches. During the years of research, developers and scientists invented many different strands of algorithms, and each one was applicable for different kinds of data.

The developed programs gained extraordinary publicity during the years when they beat humans in their proficiency, for example, 1996 when the first computer was better than the human world champion in chess. Even though Kasparow thought to recognize some "human consciousness" in the machine playing against him, it merely tested all possible moves and chose the most promising one. Significant attention also received DeepMinds AlphaGo Zero, which won against the world best Go players without having played against a human player before. AlphaGo Zero was taught by playing against itself, starting with completely random moves and learning from that experiences (Deepmind, 2019). In contrast to chess, Go has so many possible combinations of moves, that it is not possible to calculate all the movement probabilities for each turn.

Figure 2.1 (Fast and Horvitz, 2017, p. 5) shows how the reception and the presentation of AI in newspapers changes with the years. In the AI winter during the 80s, the public perception thinks about space and futuristic projects. When Deep Blue bet Kasparow in the late 90s, the term "chess" emerged in the press releases

concerned with AI. During the 90s and afterward, the topics simultaneously are getting closer to reality. Starting with "super computers" and "computer games," ending with "driverless vehicles" and "computer vision" (Fast and Horvitz, 2017, p. 5), the figure shows the range of applications connected with AI.

These achievements in games like chess and also in other, particular domains, showed one unexpected finding of AI research: Tasks that are difficult for humans are easier for AIs than tasks that are easy for humans. Things like walk, open a door, doing small-talk, or driving a car are things that can be learned by humans in a short and measurable time frame, while being world-class good at chess needs life long learning (and some talent). In contrast to this, AIs proved to be good at the latter but continuously failing at the former category (Eberl, 2018, p. 13). Car driving is only one example of many: Big car manufacturers are testing autonomous driving cars for years, but serious accidents are major throwbacks for the development (NTSB, 2019). Even though the technical community is optimistic about this topic, the big breakthrough for autonomous driving cars has not arrived yet.

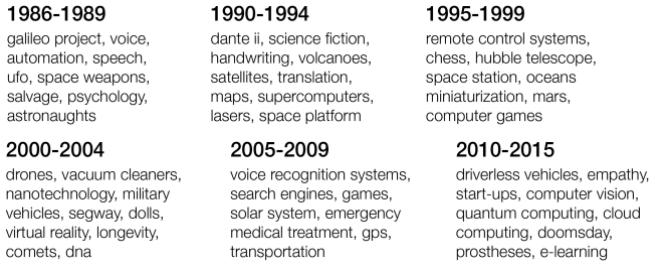

Figure 2.1: Development of Topic Connected with AI in Newspapers

For the tasks mentioned above, such as picture recognizing or language understanding, the most used techniques are Neural Networks or Neural Nets in short. With their flexible structure and variable depth, they can adapt to different kinds of problems. Interestingly, they imitate the way of working of the human brain (or brains in general) in their netlike structure. So the methods evolved from strictly mathematical and logical patterns to a more biological pattern, which maybe can be amplified up to a realistic biological model. However, we are separated several years (if not decades) from this scenario of a comparable human-sized brain emulation: With the capacity of current supercomputers, Neural Nets are just large enough to replicate a frog's brain, which is with 16 billion neurons relatively small compared to a human brain with 86 billion neurons (Eindhoven University of Technology, 2018).

Artificial Intelligences in Today's Life

AIs, like the ones described above, most often in the shape of Neural Nets, appear in various sections of our daily life. With the expansion of the Internet in all aspects, it gets easier and easier to gather data and information about people. This information then can be used with previously mentioned methods of AI. One goal, for example, is to cluster consumers, to target advertisement better on possible buyers, or to recommend similar products or services. These programs are often running in the background, unnoticed by most of the people.

Other AIs, that are also recognized as such, because they focus mainly on speech and the analysis of spoken and written words, are virtual assistants like Siri, Alexa, Bixby, Cortana or Google Assistant.[1] They can connect with different sets of services, therefore have a wide range of use cases, and are used by more and more people. For example, users can connect them to *wearables*, like headphones and watches, and *smart home devices* like smart light bulbs, the audio system, or the fridge. Use cases for text-based AIs are automatic text understanding, for example sorting and clustering of news, or the analysis of medical reports and medical papers, as Watson does, an AI by IBM (Meier, 2018).

Another kind of AI, which relies more on the interpretation of visual and sensory data, are robots, for example, for cleaning the floor or grass cutting. The cheaper and simpler ones may not seem very intelligent (when they push the same wall five times without realizing that it is solid and static), but we see a basic level of sophistication when they navigate on their own through our homes and gardens.

Even if this is just the beginning of robots in our lives, we clearly can disprove Hans Moravec's hypothesis from 1988: "I expect to see the first mass offering from the cauldron served in time for the new millennium, in the form of a general-purpose robot for the factory - and the home" (Moravec, 1988, p. 22). Also 20 years after the millennium, there are no "general-purpose robot[s]," neither in factories nor in our homes, because the scientific community had again underestimated the complexity of the "simple" task of "general-purpose."

Nevertheless, the interest in AI is just at the beginning to rise. Together with the concept of *machine learning*, which became popular after the last AI winter, those two terms built whole new fields of employment. Figure 2.2 (Google Trends, 2019) shows that today the interest in AI is higher than ever and the trends predicts even more attention in the future.

[1] These virtual assistants are distributed by the following companies: Siri by Apple, Alexa by Amazon, Bixby by Samsung, Cortana by Windows and Google Assistant by Google.

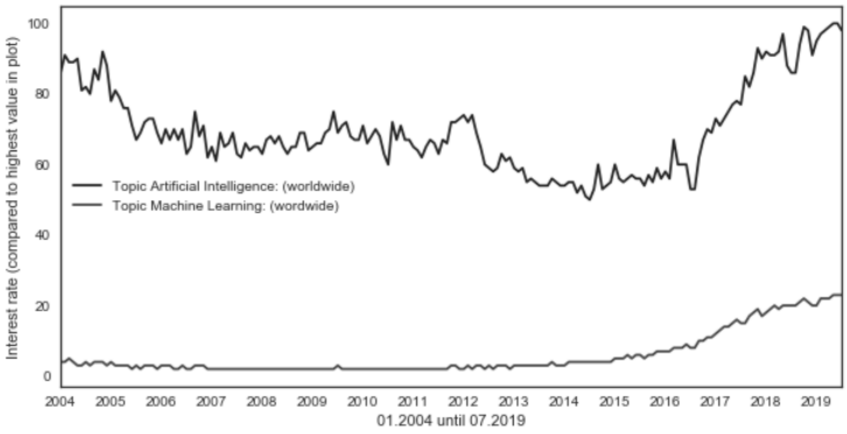

Figure 2.2: Interest Development in the Topics Artificial Intelligence and Machine Learning

Ways to Develop an AI

We have seen that today we have several systems that claim to be artificially intelligent. Companies all over the world advertise with AI in Data Analysis and other relevant sub-fields. However, what would need to happen if we wanted to achieve *real* AI in the holistic sense of the term, as described above? Is it or could it be possible to develop a *real* AI? According to Bostrom (2014) there are several ways to achieve artificial intelligence. Although his work is mainly concerned with "superintelligence," he and also other authors (Shanahan, 2015, p. 86) theorize that each emergence of true AI will automatically and immediately lead to a superintelligence because of its optimizing nature and structure.

First, Bostrom talks about artificial intelligence in the technical meaning of the term: The building of broader and more sophisticated **machine learning** models will lead someday to a model which is trained extensively and can learn new things just like a child (Bostrom, 2014, p. 42). A role model here is the natural evolution because there it was possible to create natural intelligence with a lot of trial and error. Humanity was able to recreate other natural skills, such as flying or sonar systems (Bostrom, 2014, p. 43), why should we not be able to recreate intelligence as well?

Another option would be the emergence of intelligence from **networks** or organizations, such as the Internet. Many people are using data structures, sending packages, evaluating search results as well as provided information. It is entirely

possible that from this immense collection of knowledge, AI could emerge spontaneously (Bostrom, 2014, p. 77) although Bostrom thinks that this way is rather unlikely, compared to the others.

A third option could be the **uploading** of real and existing minds into some technologically created environment. This procedure would prerequisite an emulation of the raw human brain, so to say a digital model of neurons (Bostrom, 2014, p. 51), as mentioned above. The scanned and digitized human brain would be implemented, similar to a computer program, within the neuron emulation (Shanahan, 2015, p. 15). The question is whether this method leads to intelligence. We do not know how this re-construction could function or if it would behave like a human brain at all.

It is not clear if one of these described ways will ultimately lead to AI. Some scientists even state that real intelligence in the holistic meaning of the term will never be possible. We will return to this question in the following sections.

2.2 Robots as Embodied AIs

Most often, when we think about AI, we may have no concrete picture in mind. We could think of some white code on a black screen, of green digits like in "The Matrix" or an abstract face like *Viki* in "I Robot." However, we can see some AIs because they have a physical body. Those are called robots. In his basic work about robots Jordan notes that the term "robot" is difficult to define (Jordan, 2016, p. 24), similar to the problem with the term AI, albeit it is, in general, less abstract. For his work, he uses the definition that a robot is,

> [a]ccording to pioneer roboticist George Bekey, a "machine that senses, thinks, and acts. Thus, a robot must have sensors, processing ability that emulates some aspects of cognition, and actuators." Culturally speaking, a robot tends to be a mechanical entity that exhibits human-like capabilities. (Jordan, 2016, p. 252)

The mentioned "cultural" notion can be observed often in other less general definitions, where we define robots always in comparison to humans (see, for example, (Klass, 1983, p. 171)). Initially, the word robot comes "from the Czech writer Karel Čapek's play *R.U.R.: Rossum's Universal Robots*, in which the word carried suggestions of heavy labor, even of slavery" (emphasis in original) (Seed, 2011, p. 59). Over the decades, the word was included in our standard dictionary. The origin of the term is only one example where Science Fiction literature, technology, and research overlap.

While the term *robot* only slightly brushes the connection to the concept of *humanity*, the term *humanoid* includes this relationship more explicit. Humanoid robots "basically resemble the human body shape" (Appel et al., 2016, p. 472), but the differences would be apparent: no one would confuse a humanoid robot with a real human. Therefore, all robots we currently know, that have, for example, a head and two arms, are called *humanoid robots*.

The next step towards a more human representation is called *android*. Though the term seems quite modern and new, android is indeed older than robot and was already used in the nineteenth-century (Jordan, 2016, p. 51). It describes creatures that are "capable of passing under superficial humans" (Klass, 1983, pp. 173-174) if not thorough detailed examinations and Jordan adds, it "refers to non-human beings that have flesh-like exteriors" (Jordan, 2016, p. 51).

Real androids may not be known in reality yet, because it is challenging to imitate natural behavior and the human skin surface. However, there has been some progress, for example, when we look at artificial replacements of limbs for disabled people. The prostheses are getting better and better. This combination of human and technology is called *cybernetic*. The people in question are therefore called *cyborgs*, a acronym of *cybernetic organism* (Seed, 2011, p. 62). Some people argue that a cyborg has to "consist[...] mainly of machine parts that dominate in mass but remain under the control of the natural part, 'essentially, a brain in a box'" (Haney, 2006, p. 20). The development of "a brain in a box" may still be on the agenda, but research and development are getting there step by step.

2.3 Interacting with Technology

Because technology is everywhere in our lives, many researchers are concerned with our relationship to our technological devices. Some authors of scientific studies even mention explicitly that this research aims at the future, where technology and therefore AI will inhabit our daily spaces more and more (see for example Appel et al. (2016)). In the paragraphs below, we will come across different concepts concerned with humans and technology, like the uncanny valley, technophobia, and anthropomorphism towards devices. Additionally, we will talk about stereotypes as well as consciousness, and in the end, we will examine the concept of technological singularity shortly.

The Uncanny Valley

We recall from above that we defined an android as a nearly human-looking and behaving robot. If we imagined such a robot would exist, how would humans react to such a creature? Would they accept it as equal, or would they point out the differences? Today there are scientists concerned with similar questions. There is research going on to analyze the behavior and the feelings of humans towards robots. Based on the results, the theory of an "uncanny valley" was constructed (Gray and Wegner, 2012, p. 125) to illustrate how the acceptance of humans shifts when robots in their surrounding become more human.

In general, the more human a robot is looking, the higher is the liking of humans. This acceptance rises until this moment where the robot is *too* close to a human appearance, here the acceptance drops. This drop is called the "uncanny valley." Scientists have not agreed on why this drop occurs. Gray and Wegner (2012) for example showed in their experiments that this might happen because of a "perception of mind" in the devices or robots in question. Sometimes we may think of our technical devices if they had feelings (see the section about anthropomorphism), but when we experience that this device really has feelings or shows emotions, this would be uncanny and strange (Gray and Wegner, 2012, p. 129).

Technical Anxiety / Technophobia

This *uncanniness* may develop into caution and maybe even into fear towards innovations, which is a frequent picture for the relationship between humans and technology. With each evolution in technology, for example, with the industrialization, humans feared to be replaced by machines, what in the end, at least until now, could not be confirmed (Precht, 2018, pp. 17-18.). However, this fear is rooted deep and also can be seen today when talking about innovations. This feeling concerning technological progress is called "technological anxiety" (Sims, 2013) or also "technophobia" (Seed, 2011). The last term was already coined by Isaac Asimov in the 1980s, who described the negative and pessimistic view of the society on future and technology. Where does this feeling come from? What could lead to this rejection, besides the automatic doubts regarding anything new and anything that could change our known way of living?

Sims (2013), who dedicated a whole book to this topic of tech anxiety, articulates two reasons why to be afraid. On the one hand, there is "the fear of the human body becoming obsolete" (Sims, 2013, p. 8), because specialized machines are always built to do things better, faster, with more precision in comparison to humans. Therefore, when innovations prove to be so good, it is only rational to use machines instead of humans. This replacement does not only take place for work

with force and body, but also for mental work, which can be easily automatized and optimized (Precht, 2018, pp. 118-119).

It is not enough to talk about humans in their role as workers. Besides, there is the fear that through the so-called post-humanism. We concentrate our actions more and more on our mind and less on our body. Decades ago, communication was an action that required the whole body to be physically active, for instance writing a letter, carrying it to the post office, or buying stamps and throwing it into a mailbox. Today, the process of communication is reduced to typing or recording speech messages (with even less physical movement). In the future we may even be able to send messages directly via thought. Our body loses relevance step by step, and this is accelerated in general by every technological innovation.

On the other hand, fear comes from our observation that machines and technology are becoming more and more similar to ourselves, which leads us to the question: What separates us from machines? We will come to this question in more detail in the section about self-awareness and consciousness, but it seems clear that a distinction is not as easy as is may seem on a first view. While machines are becoming more like humans, we have to think about what separates us from machines - but also where we will find similarities. Maybe machines are already more similar to us than we want them to be?

> [W]e fear [...] that they will see us the way modern technology has us seeing the world: as resources to be mastered and exploited. It says something about our worldview that we are terrified that anything non-human might treat us the way we treat everything non-human. (Sims, 2013, p. 10)

Maybe machines will also start to think just about themselves and see us as workers, slaves, producers, precisely in the same as we see trees, pigs, cars. This thought makes us feel uncomfortable because it also means that we are not driven by humanity but by the same things as machines. "We think like machines," and this proof is what we fear (Sims, 2013, p. 12).

Anthropomorphism

In contrast to us, seeing ourselves like machines, we also start to see the AIs and technical devices more and more as human-like beings. With the concept of anthropomorphism, the authors describe the behavior of transferring human traits such as emotions, thoughts, and reactions to non-human beings or things. A first clue gives Jordan when he writes about "The Anthropomorphic Effect" (Jordan, 2016, p. 211). He cites people who report different human behavior when a robot

is in the same room, for example, service robots or robots in military contexts. The described effect did not only include human behavior *towards* the robot but the robot also "changed how people interacted with one another, in largely unanticipated ways" (Jordan, 2016, p. 216).

Reeves and Nass (1996) conduct more extensive research on this subject and construct many exciting experiments with surprising results. In "The Media Equation" (Reeves and Nass, 1996) they looked at how the participants of their experiments interacted with computers, not with robots or simulated AI, but with simple, old desktop personal computers (PCs). In total, their studies confirmed the trait of humans to perceive computers as "social actors." It does not even need "virtual reality to create the sense that another person is present" (Reeves and Nass, 1996, p. 25), the people react to the computer as if it was another person. Reeves and Nass observed politeness in reaction to computer interaction, as well as the building of group identities and the positive influence of compliments and praise.

These results may sound very artificial and not realistic, especially when we remember that the participants used the computers for the first time and (nearly) only interacted via text messages. However, the results of the experiments speak for themselves. We can only imagine how these results would look like if the participants used the device longer than half an hour and with more possibilities of interaction (for example, voice and video chat).

Gender Stereotypes

This perception of technical devices as social actors does not only change our behavior with and in front of those devices, but it also changes our expectations and views on that device. Several recent studies have shown that people build constructs of "gender, age, personality" in their heads (Sandygulova and O'Hare, 2015, p. 595). This construction happens, for example, when they are interacting via voice with robots. Depending on those mental constructs, we also expect specific behavior of the device (Reeves and Nass, 1996, p. 177). Researchers call those indicators that are used to build those mental expectations "social cues" (Feine, 2019). It means that we as humans connect several outward visible indicators (like for example gender, age, voice frequency, word choice, answering speed) with personality attributes, which we expect from the person in question. Especially when we encounter a technological device via voice - even if it a synthesized, computer-generated voice (Reeves and Nass, 1996, p. 168) - social indicators are attributed, amongst others also gender. Moreover, because of that, "people may have predispositions toward a robots knowledge base or usefulness based on its perceived gender" (Rea et al., 2015, p. 556). It even creates confusion if the technological device does not apply to those stereotypes (Reeves and Nass, 1996, p. 177).

Especially virtual assistants, which interact with humans directly and quite frequently, are provided with a friendly name, a voice and sometimes even an avatar. Most often, those features lead to an implicit attribution of gender; sometimes, this happens also explicitly. The reasons why a particular gender is attributed to virtual assistants or also to conversational agents like, for example, chatbots are not always clear. On the one hand, we have *Cortana* with a female voice and based on a female computer game character, *Alexa*, which is explicitly described as female, *Siri*, *Bixby*, and *Google Assistant*, which at least come with a female voice as the default setting (Steele, 2018). They all represent friendly, supportive, and never offensive characters, which should help in daily life tasks (Ramge, 2018, p. 16).

Then, on the other hand, we have AIs such as *Watson*, which works as a medical assistant and competes with quiz champions in Jeopardy. Watson has a male voice (PBS NewsHour, 2011) and a name reference to the most known Watson, the friend, and assistant of Sherlock Holmes. When the creators of virtual assistants were asked why they used female characters per default, they give profit calculations based on the compliance to user expectations as reason (Steele, 2018). They tested the achievement of different options and got the best results for those female voices. Also, users probably expect those virtual characters to be female, because they serve traditional female tasks like scheduling meetings, reminding of tasks, giving information and advice.

Researchers from the Karlsruhe Institute of Technology (KIT) launched a kit to select the right characteristics for a chatbot based on exactly those expectations (Feine, 2019). They collected the scientific reports of social cues and which influences where reported. For gender, one of their sources reports that "the female character may be more effective than the male character because of the existing association between female gender and qualities such as empathy" (Hone, 2006, p. 242). Figure 2.3 (Feine, 2019) shows all social cues they developed based on research results. Some cues indicate concrete expectations the user may formulate towards the conversational agent, and others reinforce certain stereotypes connected to the cue.

Studies show that stereotypes or mechanisms that are used to evaluate humans without further information are also applied to robots (Reeves and Nass, 1996, p. 163). We transfer experiences from human-to-human interaction to human-to-robot interaction, which leads to impressive experimental results: Reeves and Nass found out that experiment participants rated a computer with a female voice better at giving tips about relationships and love and worse at giving technical information about computer, compared to a computer with a male voice (Reeves and Nass, 1996, p. 165). We internalize these concepts via socialization; even small kids refer to them. When a study let children draw pictures of robots, the

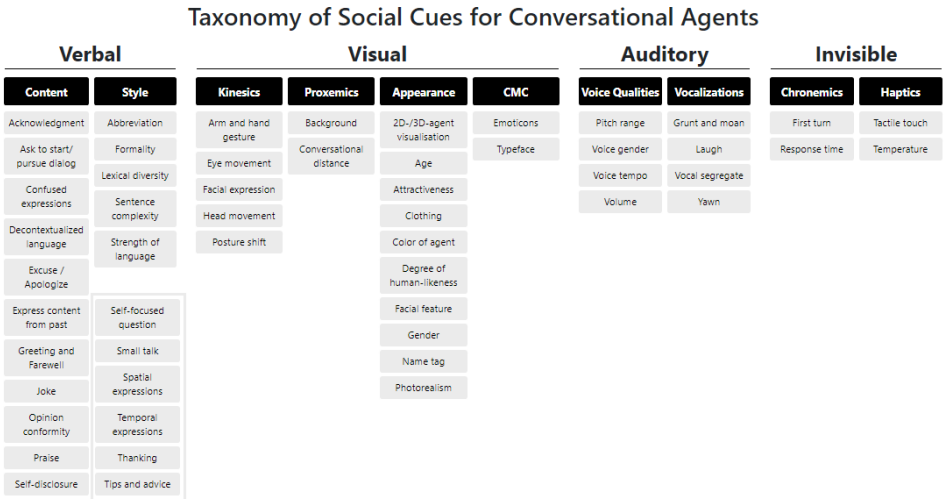

Figure 2.3: Social Cues Developed by Social Cue Design KIT

results showed that in this experiment, female robots were "associated with positive characteristics such as friendliness" (Obaid et al., 2015, p. 504). This association is especially interesting as male robots had no such unique associations; the children perceived them as neutral. This *male normativity* can also be seen in other contexts (Reeves and Nass, 1996, p. 168) and we will come in chapter 4 back to this topic.

Self Awareness and Consciousness

When humans are interacting with AIs, they do not only apply internalized stereotypes. They also expect other human traits based on the AI's intelligence. Even though intelligence can vary on different levels, with higher levels of intelligence, we as human beings also assume some reflective behavior. The more difficult problems get, the more solutions there are, and the less accurate our success measurements are - the more our deciding system needs to reflect itself and its situation to consider possibly fitting solutions. We implicitly expect this from intelligent systems, even though it is not real yet (as far as we know). This procedure of reflection is called *self awareness*. It is one of two concepts which divide our human understanding of intelligence (the holistic approach) from the technical understanding.

The second concept can be explained again with the help of the dictionary definition from above. "Understand language" was one of the example applications mentioned in the definition of AI. What does *understanding* mean? The technical

approach would argue that a text processing system can indeed understand phrases and sentences and give - automatically - the right answer or information. The holistic approach would question this. Just because a system calculates the most probable or most frequent sequence of words as an answer to another sequence of words (in a given situation), it does not mean that this system understands the *meaning* of those words and sentences. This process is described by the term "consciousness which is the state of understanding and realizing something" (CED: Consciousness, 2019).

The literature gives different perspectives on mutual understanding: Either we need a "shared assumption of mutual empathy" (Shanahan, 2015, p. 113), which means that we need to be able to imagine the self-awareness of the other; or we "assume consciousness of the other person" (Hubbard, 2011, p. 443). For both assumptions, we are not sure that they hold. You as a human being can just *assume* that other humans are conscious because you know that *you are*.

When we want to talk about conscious computers, we could do an intermediate step and talk about animals. The current research showed that certain species of animals possess abilities that could imply higher intelligence than we have ever thought (Hunt, 2017). Are they conscious? Do they have complex feelings? "If an animal act as I do when I am afraid, is it not reasonable to call its mental state fear?" (Moravec, 1988, p. 39). Is this just the result of our anthropomorphic behavior? If a robot would behave like it was afraid, why don't we call it a feeling and name it fear as well? Just because it is a mathematically calculated reaction and no chemical, *human-like* fear?

Some scientists claim that consciousness never can be reached by artificial or mechanical constructs (for example Haney (2006) or Rojas in (Dorn and Rojas, 2018)). They argue that mind and body are tightly connected, which means that the mind cannot exist in its full shape without its body. According to them, our awareness of being, our sense of being alive, is only possible through the feeling of our body. This feeling happens through sensual stimulation or the transportation and reception of chemical information within the human brain. Those criteria maybe can be simulated by computers, but they can never truly replicated it. This definition then also implies that there will never be any other awareness than *human* awareness. Maybe this view on awareness is too narrow. It sometimes seems that scientists use those two concepts of self-awareness and consciousness to establish borders between *us* and *them*. However, as technology gets more and more intelligent, and at some point eventually conscious, we should think about how we would react to conscious non-human beings and which laws and rights had to be applied to them (Hubbard, 2011).

Singularity

The biggest and maybe most ultimate view on the relationship between humans and technology is the *technological singularity*. The technological singularity, also set equal with *superintelligence* or *intelligence explosion* by Bostrom (2014), generally describes the moment when "exponential technological progress brought about such dramatic change that human affairs as we understand them today came to an end" (Shanahan, 2015, p. 11). This change can be understood in two distinct ways: First, and this is the optimistic view, human affairs will end because the superintelligence takes care of everything that a human could be concerned with. It would be a utopia, where no one needed to work or to worry about food, living conditions, or anything else. Second, when we look at it more pessimistically, it could mean the ending of human affairs simply because there are no humans anymore. This can have several reasons, for example, the technology thinks of us in terms of resources and simply uses or consumes us until nothing is left (O'Connell, 2017, p. 83), or it is programmed to make humans happy and discovers that dead humans are the most efficient and happiest humans because they do not have any wishes (Precht, 2018, pp. 182-183).

Chapter 3

The Fictional View on AI

It is understandably complicated to examine AIs, their influence, how they will change our relationship to technology. In the previous chapter, we have seen that AIs are still in their infancy and therefore, miles away from being autonomous or on a human-level of general intelligence and consciousness. We do not even know in detail what it needs to be *truly* intelligent. But, and the previous chapter has proved this, the topic is already important for our everyday life. We generally see technological innovations very skeptical, despite fears and cautiousness, we are willing to let AIs and the belonging technology in our most private retreats. In the future, this topic will become more and more relevant (Ruge, 2012, p. 59).

Albeit there are studies investigating human-robot-interaction in real life and with real robots (see for example Sandygulova and O'Hare (2015) for perceptions of children or (Alves-Oliveira et al., 2015) for the work with older people), those studies are experiments only. They test interaction in an artificial setting and a minor scope. We are not talking here about the interaction of *society* and AI. Additionally, we are only talking about robots, not about invisible AIs, which are not advanced enough to be tested in real-life experiments. Indeed, there are studies which simulate the presence of a generally intelligent AI, or at least a more intelligent AI as those existing now, while the AI is simulated via chat and controlled in the background by a human being. In these experiments, the "experimental participants are given the impression that they are interacting with a program that understands English as well as another human would" (Kelley, 1983, p. 193). This type of experiment is called *wizard of oz* experiment.

Again, there is no real, no large scale interaction visible. Other scientific authors came to a similar conclusion: "Given the limited nature of the current state of development of artifacts [...], a useful source [...] for artifacts is science fiction" (Hubbard, 2011, p. 455). We can take Science Fiction stories and examine *them*

to get a clue about which reactions and situations to expect in real life. Hubbard, for example, takes fictional AIs into account to analyze their status of personhood according to his definition (Hubbard, 2011). Other studies focus on SF to conduct in-depth research, for example, on technological anxiety (Sims, 2013) or the development of consciousness (Haney, 2006). There are also studies which concentrate on movies to say something about how we should anticipate the rapid development of robots and AI and which reactions we can expect (Ruge, 2012).

Two assumptions are underlying these kinds of studies: First, SF is trying to set a realistic view of the future (or on a parallel world to today) with the inclusion of robots and AI. Second, those stories have some impact on society, or at least on those people reading, watching, interacting with those stories. We will see in the following chapter to what extent these assumptions hold and what this means for the paper in the end. But first, the chapter provides an introduction to SF, what it is, what it is about, and why it could be rewarding to analyze it for the real-life view into the future.

3.1 SF - Aspects of the Definition

Science Fiction (SF) is a term which is so broad that it is difficult to define. In general, SF describes a sub-part of fiction, which uses elements of science. Science, in this case, can be the use of new technologies or the view on new developments. The story is often set in the future, but it does not have to be. As many understandings of SF exist in general, as broad are the approaches obtained from the literature to describe SF. We will analyze some of these approaches according to the underlying question of this paper.

The very first association would be "Science Fiction is a genre," since the books in a library are sorted by its genre and SF is one of them. Sometimes SF is not classified as a genre but as a sub-genre together with Fantasy as *Speculative Fiction*. SF and Fantasy are indeed closely related in their way of telling stories, but they need to be separated in terms of content. While Fantasy uses elements of magic, inexplicable powers or events and worlds that are not related to our real world, SF uses none of these elements and in general relies instead on science as an explanatory factor. Of course, given the related nature of those two terms, the boundaries are blurry, and some works fail to be classified as either-or.

However, because of the variety of topics inside SF, it also fails to be classified as genre on its own. Ruge (2012) refers to a definition that describes a genre as a homogeneous canon of topics and motives, which does not match with SF. When the works collected under the SF term are so entirely different, how can we conclude that something is SF and something else is not?

According to most of the authors trying to define SF, the direct or indirect connection of the story and the world described to our known world is an important aspect. Seed puts this in the following sentence: "It is helpful to think of an SF narrative as an embodied thought experiment whereby aspects of our familiar reality are transformed or suspended" (Seed, 2011, p. 2). The author turns the world into something new but not wholly new. Ruge calls this the "continuity" of SF, the assumption that, at least in theory, you could draw a direct line from our reality to the story (Ruge, 2012, p. 33).

The second aspect of SF is the "science" part. Attebery (2002) focuses on this aspect in his definition: "Science fiction is a system for generating and interpreting narratives that reflect insights derived from, technological offshoots of, and attitudes towards science" (Attebery, 2002, p. 2). So because our reality is included in each SF story, the focus also lies on the differences between the two worlds, *because* of the scientific evolution involved.

When we look closer at the mentioned definitions, we find a third aspect that is special in comparison to other genres or topics: SF is often called a "narrative" (Seed, 2011, p. 2) or "a system for generating [...] narratives" (Attebery, 2002, p. 2). The term suggests to look on the *way* of storytelling and not on the content or on the story itself (Ruge, 2012). This concept of the narrative is important to remember, as it divides critical cases of SF and Fantasy clearly: Regardless of the world in the story, it depends on how the story explains the world, how it credibly describes the connectivity to our world (Ruge, 2012).

For the following chapters and especially the following sections the slightly extended definition of Shaviro (2016) will be the underlying assumption: Science Fiction is a narrative "that goes beyond our present scientific knowledge and technical abilities, but that - at least in principle - remains within the bounds of scientific possibility" (Shaviro, 2016, p. 22). This definition combines the aspects mentioned above of continuity, relation to science, and view as a narrative in a short and understandable way.

3.2 Relationships between SF and the Society

Ongoing from our definition about SF, we know now that there is a tight connection between the real world and the SF worlds described within stories. We can examine this connection from different perspectives. These include, among other things, the relationship between SF and science, the role of SF as a mirror and as a critic of society, and the influence of feminist SF. Those perspectives are explained one by one in the following.

Science and Science Fiction

Science is one of the most critical connections between SF and our society. Not only because it is a defining topic as we have seen in the section above, but because of two reasons that make the connection even tighter. On the one hand, we observe the influence of actual research on the topics covered in the SF stories (Drenkpohl, 2009, p. 57). This influence may occur because the topic is in the news or the writers are personally interested in the developments in this particular research area. Often, there is some professional motivation behind this interest as studies show that SF authors tend to have a professional career background in natural science before they started their writing career (see for example (Appel et al., 2016, p. 472) or (Bainbridge, 1982, p. 2081)).

On the other hand, "science fiction motivated individuals to delve into a particular problem or to pursue a scientific career in the first place" (Appel et al., 2016, p. 472). Their stories influenced the readers to get involved with science, maybe even with the concrete topics mentioned in the stories. This activation effect is also and especially true for the AI topic, because "some AI researchers and AI enthusiasts are indeed sufficiently open-minded to read SF and be intrigued, simulated, even influenced by its extrapolations" (Barnouw, 1983, p. 336). This strong connection may be changing because authors like Bainbridge (1982) fear that the emerging number of women within the authorship of SF stories would weaken the influence of science. Women tend to come from different backgrounds than natural science before writing SF and also tend to write less hard SF.[1]

Talking about *the Other*

Because SF is always related to real life, it focuses mainly on the differences of the story world compared to our reality, or to the reality in which the story was written. The difference manifests the construction of *the other*. The others are mainly those people, things, machines, aliens that are different from our experience, different from *ourselves*. Attebery calls this "the pull toward strangeness," which can be used in SF better than, for example, in non-speculative genres and "conventional realism" (Attebery, 2002, p. 5).

Suvin (1979) calls this phenomenon "estrangement," which refers exactly on that transformation of something known into something possibly unknown. In his understanding, the "normative system" (Suvin, 1979, p. 6), thus the notion of the norm and normality, is turned into something new - "the novum" (Suvin, 1979, p. 70) - by either "extrapolation" or "analogy" (Suvin, 1979, p. 75). This novum is

[1] *Hard SF* is SF that deals with natural science as a significant topic, while *soft SF* rather covers content from social sciences.

not only relevant in consideration of the reality of the author but simultaneously "validated within [...] its interaction with reader expectations" (Suvin, 1979, p. 80).

Based on the novum we can think and read about questions and their possible answers that can only be asked in SF, for example: "What is it like to be an computer?" (Barnouw, 1983, p. 333). Fiction can provide a deeper view into specific situations or people as ordinary descriptions because the reader identifies with the point of storytelling. The character telling the story shares its experiences, and these are turning into the experiences of the reader itself (Appel et al., 2016, p. 473), or as Barnouw puts it: "Fiction has always tried to penetrate the insulating screen of subjectivity, it has a vocation for imagining the Other, indeed, it has thrived on examining what it is like to be the Other" (Barnouw, 1983, p. 333).

Looking into the Mirror

The tight connection of SF stories and our real-life forces us readers not only to enjoy the story but also think about our own lives and situations. Even though we are reading about other planets, universes, time frames, in reality, we are reading about ourselves. This self-centredness means that each SF story is always a mirror for the societal reality in which the story was written (Ruge, 2012, p. 7). It is a mirror "in which they see themselves reflected, and through which they come to understand the fantastic shapes of their own identity" (Haslam, 2015, p. 1).

It already begins with the childhood: Identity, that means the feeling of who we are, is always defined by the knowledge of who we are *not*: the others. We always define our self over the differences to others, in various nuances. So SF is the perfect narrative to explore this "artificiality, simulation and the constructed 'otherness' of identity" (Janes, 2007, p. 92).

According to that, when we read about others, let them be aliens, AIs or robots, we think about what separates us from them, what defines us for example as human and those as something different. We see ourselves in a mirror, from the outside, with different eyes. This "affect[s] how we feel, think, and act", as recent studies have shown (Appel et al., 2016, p. 477).

SF as Critique

Of course, by thinking about ourselves, we do not only register the status quo, but we also see ourselves with more distance. We realize, for example, what is wrong with our society or where is the difference compared to other, subjectively maybe better societies. Suvin describes this as an "approach tending toward a dynamic transformation rather than toward a static mirroring" (Suvin, 1979, p. 10).

One the one hand we can think about *ideal future worlds*, which is the so-called utopian part of SF, one the other hand "sf can help us to grapple with some of the conundrums of our present world by working through their possible conclusions and outcomes; but then sf has always been more concerned with critiquing the present than with constructing the future" (Mitchell, 2006, p. 125). It may seem odd as SF is known in general for *futuristic* stories, but in fact, we can observe that the stories tell more about today than about tomorrow. Suvin agrees to that when he writes that "the value of all SF [...] is to be found in its analogical reference to the author's present rather than in predictions" (Suvin, 1979, p. 78). Alternatively, as Seed puts it:

> The crudest reading of an SF novel is to ask 'did Arthur C. Clarke get it wrong?' Science fiction is about the writer's present in the sense that any historical moment will include its own set of expectations and perceived tendencies. (Seed, 2011, pp. 1-2)

However, this can lead to reproduction in circles: Even with the freest and fantastic literature, we are not able to escape the predetermined areas of our thinking behavior (Haslam, 2015, p. 212). We are just reproducing what we already know, only in different shades and points of view.

Feminist SF

We have seen that "science fiction is a form of storytelling that invites us to challenge standard notions of nature and culture" (Attebery, 2002, p. 4). Because of this invitation, SF was, since the early times, also a medium for women to express their opinions. The topics range from suppression and power to their points of view on a male-dominated world. Seed reports that the first utopia by a woman was written in 1666 by Margaret Cavendish (Seed, 2011, p. 90). Based on this idea, many other women followed her.

A notable increase in the number of women utopias - most of them feminist utopias - was observed during and after the 1970s. In this time, higher awareness about power and suppression was noticeable through all marginalized groups in the United States. Writers like Alice B. Sheldon, Joanna Russ or Ursula Le Guin (to mention a few of them) were very famous (Seed, 2011, p. 90-92). They used their popularity to make the female view on society structures and living together accessible to a broader audience than before. In the tradition of these writers and thinkers, in 1991 an SF literary award was founded and named after Sheldon's male pseudonym James Tiptree, Jr. (James Tiptree Jr. Literary Council, 2019). This award is given each year since then to a story "that explores and expands gender roles" (Attebery, 2002, p. 6).

3.3 Fundamentals of Fictional AIs

We have seen in the previous section that SF stories have a long and deep tradition within human history. Therefore, to understand recent SF stories about AIs, we must be aware of the long-lasting dream of humanity to create artificial life. Artificial life, which is the creation of life without the help of nature (Jestram, 2000, p. 11), has been a topic since ancient times. As recent stories often include references to characters or situations from older stories, the section below will explain some of those stories.

Myths and Legends

The first person handed down in Greek mythology to have created life without the consent of the gods (even if he was a son of the gods himself) was *Prometheus*. According to legends, he formed the first humans from clay and gave them fire. For this action, godfather Zeus punished him. Other creatures were built by *Hephaestus*, the Greek god of blacksmiths. One of them, namely Talos, built from metal and bronze, guarded the island Kreta and was able to beget a son (Jestram, 2000, 12). With the help of gods, also humans were able to create living creatures, for example, the king *Pygmalion*. Because no woman pleased him, he ordered to build a statue of the prettiest of all women. He fell in love with the statue and because Aphrodite sympathized with him, she gave life to the statue (Jestram, 2000, 13).

It seems clear that in all those early stories, the humans were not able to create *complete* life on their own. Either the creations were animated by some divine gesture or magical moment, or the creations stayed incomplete, rough, and puppet-like. A common theme in those stories is the *hubris* of the creators, where their pride and arrogance, especially against the directions of gods, lead them into personal misfortune. Because of the explanation based on magic or divine power, and not on science, those stories are therefore myths, and not SF as we understand it today.

An essential step towards this understanding of science took the medieval alchemists with their experiments. In their striving towards non-ending life, they also explored more about the creation of life and the first automatons (Jestram, 2000, 17). Similar to the Europeans trying to create small "homunculi," we accordingly find hints in Jewish (Golem) as well as in Arabic (Takwin) traditions during this time. The more understandable the world became through science and research, the less magic is needed to animate objects. We can trace this development in the following section about the first real SF stories.

Early Science Fiction

Mary Shelley's *Frankenstein* (1818) is regarded as the first SF novel, as the main character Doctor Frankenstein rejects alchemy and creates his monster based only on scientific decisions. The alternative title of the novel *The modern Prometheus* links back to the already mentioned story of Prometheus, who created humans from clay. In this modern version no clay is needed but some other tissue. Some sources report that Frankenstein used the body parts of dead humans, but the novel does not mention this explicitly. This tissue, whatever it is, is then transformed into an animated being. The creature moves and reacts, it even speaks and thinks, and could, therefore, be regarded as the first artificial intelligence in the newer history of literature (Shelley, 1994).

Samuel Butler made the next important step with his novel *Erewhon*. The title is the word "nowhere" spelled nearly backward and refers to a country that should resemble the homeland of Butler so that he could formulate societal criticism within his fiction novel. Erewhon also includes three chapters about machines and their evolution. The speaker within the novel thinks about conscious machines in comparison to conscious animals. It is especially impressive when we keep in mind that he wrote this text in 1872:

> "Reflect upon the extraordinary advance which machines have made during the last few hundred years, and note how slowly the animal and vegetable kingdoms are advancing. [...] May not the world last twenty million years longer? If so, what will they not in the end become?" (Butler, 1872, chap. 23)

Another milestone in establishing the trope of artificial humans was "The Future Eve" or "Tomorrow's Eve" by Philippe Auguste Villiers de l'Isle-Adam. This story about a fictional Thomas Edison and his friend suffering from lovesickness made the term *android* popular. Again, and comparable to the kind Pygmalion, a perfect wife was made (de Fren, 2009). The story focuses mainly on the creator and the creation process, how this miracle was scientifically feasible.

20th Century

As mentioned before, the Czech author Karel Čapek was the first to use *robot* in today's sense of the term. The play from where the term was derived is about a factory where humanoid robots replaced all human workers. They are producing more and more robots, based on artificial tissue and human anatomy, but with reduced complexity. They do not have feelings or an understanding of death and fear, and therefore, no soul. In our understanding, we would instead call them

androids because their appearance is very human-like: They even have replicated skin, so the imitation is perfect. During the story, the robots refuse to continue to work in the factory, and revolt against the owner and later against humanity in total (Čapek and Wyllie, 1920).

Only five years later the German writer Thea von Harbou invented a future city called *Metropolis*. In her city, there is a similar division between workers and owners compared to the world in Čapek's work, only that the workers are still human. The story evolves around Freder, the son of the most important man in the city, and his beloved Maria, a girl caring about the worker's children. A scientist captures Maria and transfers her appearance onto a mechanical body, which he can direct as he wishes. With the imitation of Maria, which is very realistic, Freder and his father are played off against each other and later, the workers start a revolt (Harbou, 1925).

In the 1940s, a famous author was introduced to the public; his stories characterize our views on robots until today: Isaac Asimov. In his uncountable short stories and books, he developed a future where humans and robots live side by side in equality and peace. Maybe his most famous legacy is the "Three Laws of Robotics," which were adapted and used in various books and movies afterward:

> 1) A robot may not injure a human being or, through inaction, allow a human being to come to harm.
>
> 2) A robot must obey any orders given to it by human beings, except where such orders would conflict with the First Law.
>
> 3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Law. (Seed, 2011, p. 61)

Asimov writes in a time that is later called "the golden age" of SF (Seed, 2011, p. 50) because the prospects in the future were mainly positive and humankind thought that it would be genius enough to solve all environmental and societal problems arising in the future. Therefore, Asimov's view on the creation of AIs is mainly positive. Robots are exceptionally friendly and cannot pose a threat to humanity.

This changes with the years. A very prominent example of how the AI or the computer can become an enemy of humanity (or specific humans in this case) is *HAL* from the novel *2001: A Space Odyssey*. HAL, originally built and trained to help the crew of a space ship on their way to Jupiter, began to kill the crew members one by one because it wanted to improve the chances for their journey to be successful. In its calculations, the humans on board were only negative factors (mainly because they wanted to shut HAL down), so it decided the mission was better off without them (Clarke, 2016). With this novel, some argue that the

view on computers as all-knowing, invincible and flawless intelligences changed to a more pessimistic view, but as we have seen, this view existed already before HAL. Nevertheless, and mostly thanks to the well known cinematic production by Stanley Kubrick, HAL is still one of the best known fictional AIs.

Chapter 4

Formulating Expectations on AIs

The success and the popularity of SF books in general, and about AIs in particular, are undeniable. Therefore also the scientific community is interested in this topic. Up to now, many literary studies have examined SF stories and came up with various topics around AI and its influences on people and society. The following chapter will have a closer look at three of those topics, in connection with a formulation of expectations for our underlying research question. The selected topics are concerned with the roles AIs fulfill in the stories, how their characters are constructed, how they are living together with humans, and how these aspects change over time.

4.1 Gender Roles and Their Manifestations

The first topic that will be under investigation is the topic of gender. Thereby not only the gender of AIs will be of interest, but also the gender of characters in the story interacting with the AIs and the AIs role in context those other characters. As already mentioned in the chapters above, there is no real need to assign gender to AIs. It is especially true as we are talking now about fictional AIs which have even fewer constraints from the outside than real AIs. In theory, they do not need to fulfill the expectations of paying users, nor do they need to comply with market-driven decisions. However, and we have seen this above, people live better when they can directly evaluate the things they see, so we find gender assignments even for fictional AIs, independent of whether those evaluations are accurate or not.

Above the concept of *male normativity* was already mentioned, which means that our way of thinking through our socialization assumes male as standard and female as exception of that standard. We have seen this in the reactions of children who should draw their favored robot teacher, we can see it in different languages,

where the male word form is the standard. We also see it in the literature: Attebery (2002) finds this behavior with cyborg or robot bodies, that they "would have seemed to most readers to have no gender at all" when they were "marked as masculine" (Attebery, 2002, p. 6). In a male-normative world, if we think of robots, cyborgs, android beings in general, they are generally speaking more likely to be perceived as male. When we extend this observation to SF stories, we should find similar results.

Hypothesis 1.1: The AI is rather presented as a male than as a female.

This hypothesis will be the first under investigation. We should also be able to observe the phenomenon if the gender is not mentioned at all, or if the gender is definitively neither male nor female. Even in those cases, the characters interacting with the AI should in tendency, see it more as a male construct than as a female construct.

But *if* the author thinks about the gender of the character, the picture should change. It is comparable to the children drawing robots. When they think about gender and mention it explicitly, those robots are in tendency more often female. If something deviates from the norm, then there is less tendency for it to be perceived as male. It leads to the second hypothesis, which has two separate statement parts.

Hypothesis 1.2: (a) When the topic of the gender of an AI is explicitly mentioned, it is more probably female than male. (b) When gender is only mentioned and not discussed more deeply, it is more probably male than female.

In the overview of first stories about AIs, we have seen that besides the gender, the creator is an essential topic in those stories. The relationship between creator and creation is most often an extraordinary one, either because of love and possession or because the creation stood up against its creator. We only need to think of Frankenstein and his monster mentioned above.

In the next groupg of hypotheses we will come back to this trope and have a closer look at the power balance between creator and AI, but in the meantime, we will formulate expectations about the gender of the creator and the gender of the creation. With the older stories in mind, we can already say something about the gender distribution between the two parts:

Hypothesis 1.3: When the creator is male, the created AI is more probably female than male.

The creator tends to be male because it seems to be an intrinsic male dream to create (artificial) life. The literature explains this with the inability of men to give birth naturally. Women, according to Jestram (2000), have always been tighter connected to nature than men, which resulted in a general alienation from nature. While women were able to give birth to children without men, men always had a passive role in this act of creation. Therefore the literature figured this out as

subconscious wish, to create life without the aid of nature or women (Jestram, 2000, p. 21).

On the other hand, this creation tends to be female, especially when the creator is male. This situation can, for example, result from the wish to create a perfect wife, as the chapter above mentioned it for Pygmalion. Alternatively, the creator wanted to create a *childlike* creature, where a daughter seems to be less dangerous for fathers: "Unlike sons, who are potential rivals, daughters can surpass their Daddies without doing too much damage to male egos" (Attebery, 2002, p. 89). This thought is not necessarily true for real relationships, but it is a pattern frequently found in stories and tales.

For the last hypothesis within the gender topic, we quickly find examples in the real world. We have already seen that cloud voice assistant systems in our reality tend to be female, also because of economic reasons and gender expectations. Femininity is often connected to assistance, helpfulness, friendliness, and therefore, it is the right choice for those systems to make them female. In contrast, physical attributes, for example, muscles, power, physical force, are often connected with masculinity.

Hypothesis 1.4: (a) Female AIs tend to be more in cloud form, (b) male AIs tend to be more in physical form.

Drenkpohl (2009) finds this, for example for cyborgs, while the cyberspace as the non-material world, where force and muscles do not have any impact, is rather regarded as female (Drenkpohl, 2009, pp. 159-160). Also Sandygulova and O'Hare (2015) supports this when he writes that those "three characteristics [are] rarely seen in combination: femininity, power and artifice" (Sandygulova and O'Hare, 2015, p. 6). Do we find this expectation also in our data? Are female AIs rather working in the background, without having a forceful body to support their interests, while male AIs have it?

4.2 Relationships of Power and Control

The second topic deals with different aspects of *power*. Power, in general, has always been an essential topic in fictional literature, if not *the* topic overall. Especially when it comes to the process of creation, the creator exercises power over the creation, because it then exists just because of the will and control of the creator. The scientific literature provides different perspectives on this topic.

Just as the term robot suggests in its translation into English (see chapter 2), we talk about workers when we talk about robots. In some translations, even the term "slave" or "forced worker" is used instead of "worker" (Jordan, 2016, p. 30). The robots were built by humans to do their work, to replace human workers more

efficiently, either to make the life of its builder easier or to lower the number of flaws in the work. They are built as "enhancements for their makers," "as virtual proxies for them, doing work that they should more rightly do" (La Grandeur, 2011, p. 240). A robot, in contrast to a human worker, never gets tired, never makes mistakes, never complains about orders or working conditions. The scientific literature generally agrees on the notion that "the robot may be a servant [...] but apparently never an employee" (Klass, 1983, p. 175), because it can't - it mustn't - decide for itself. Ruge (2012) calls this pattern in his extensive research about robots in movies the "Golem" figure (Ruge, 2012, p. 65) (see section 3.3).

Slave, servant, worker, all connotations of this relationship imply a superior *master* figure, that is most probably human, and a subordinate figure, the robot. The relationship is constructed like that because the human-built the robot and maybe even added some restrictions that the robot never can harm a human (see laws of Asimov in section 3.3), never can get away, and never be truly free. It is a pure power-dominated relationship between superior and inferior. This relationship that may also reflect the relationship between humanity and technology in total builds the first hypothesis within this topic:

Hypothesis 2.1: The AI appears as a servant for human beings.

But La Grandeur (2011), cited already above, writes that AIs are "too aware of their own power to be good servants" (La Grandeur, 2011, p. 242). When a servant or a slave knows about its power and knows about the fact that it is stronger than its master, nothing will keep it in this unnatural master-slave relationship. As the robot is constructed to be always better than the human, smarter, faster, stronger, with fewer flaws, the robot will try to succeed in its attempt to escape the human authority.

And the human, on the other hand, the creator of those AIs, underestimates the power of the creation. Maybe the human doesn't want to see the flaws, the weaknesses of the inventions. Perhaps the human is even too indulgent so that the expectation to be overthrown by the own creation is far away. This expectancy that the creation will harm its creator is called "Frankenstein Complex" (Seed, 2011, p. 61) in the scientific literature, because on the one hand, Frankenstein's Monster exists only through and because of Frankenstein, on the other hand, it will never settle in peace when the creator is still alive (and maybe still has control over it).

In the previous chapter and through reference to Frankenstein, we have seen that this is not a new topic. Humans have feared new things in technology as long as we can think, and the more powerful the machines get, the more dangerous they possibly get. With the development of AIs, this problem gets more relevant as they are constructed to be in each way better than humans. And because of their consciousness and intelligence, they can update and optimize themselves. Of course, merely being better does not make something or someone dangerous by

definition. But our feeling tells us that AIs may be hazardous, and stories adapt to our fears and extrapolate them. The next hypothesis tests the portrayal of this feeling:

Hypothesis 2.2: The AI is dangerous for (a) human beings in general, (b) its owner/creator, (c) humanity in total.

This hypothesis is divided into three sub-hypotheses because the scientific literature mentions different kinds of *danger* depending on who is endangered. Are humans in general in danger? Or do we find, in alliance to the Frankenstein Complex, a fixation on the creator (or the owner) of the AI? Or do we find danger for humanity in total? This hypothesis implies that the AI is dangerous in its own accord, and not because its owner said it so. It is the decision of the AI itself to harm humans, and not because of programming or equal. If we think about an AI programmed to kill, we should have a look at the next hypothesis.

Because when we think about fields where AI and robots are applied today, we quickly come up with military drones, robots to secure minefields, or cloud surveillance programs. The last step in this development would be a *warrior AI*, that means an AI or a robot which was built to fight in place for humans, maybe against other humans or other robots. It is a crucial topic, as this might be a deep routing fear of humanity: The emotionless fighter out of control coming to destroy everything without mercy. It is also fascinating, as the developments toward this image we have already in our heads today, do indeed violate the laws of Asimov (Idier, 2000, p. 265), as they make it possible to harm humans. Nearly every computer can be already used to do so, and with military equipment, the case is even more evident. Are the AIs in SF stories imagined as military robots?

Hypothesis 2.3: The AI is produced and used as a warrior.

In the scientific literature, we also find a different perspective on depicting AIs. Ruge (2012) as well as (Irsigler and Orth, 2018), even though they are conducting their research on movies instead of books, see a notable distinction in body AIs and hyper AIs, which both include different notions of power. The first, the *body AI*, is regarded to be more human-like, more as a child who wants to emancipate itself from its surroundings which exercise power over it. It thinks of itself as human but learns that is is something different. As it is not clear how to live as AI under humans, it tries to find a way which often leads to conflictual situations.

The other, the *hyper AI*, has no distinct body but is nevertheless in the position to manipulate, to guide, to change humanity to its favor. It is often not bound to any physical location but has the power over a whole data net to get its information. The scientific literature tends to find this last one more dangerous than the first one. There are even tendencies to see this AI into a Christ-like figure (Williams, 1994). Do we find their pattern also in our selection of books?

Hypothesis 2.4: The AI is either presented as body AI or as hyper AI.

All these mentioned hypotheses will be investigated to get a picture of the tendencies within SF stories. To take up the possible results and insights from the first section about gender, all those hypothesis will be tested for reflections of stereotypes in the gender distribution: We have seen for example that femininity, in general, is connected to helpfulness, so do we find female AIs more in a servant role than male AIs (hypothesis 2.1)? Accordingly, do we find the pattern that stories depict female AIs less often as dangerous (hypothesis 2.2)? Are female AIs less frequently described as warrior AIs (hypothesis 2.3)? And finally, do we find that body AIs are more depicted as male, while hyper AIs are in tendency more female (hypothesis 2.4, with the assumption of hypothesis 1.4)?

4.3 Variation over Time

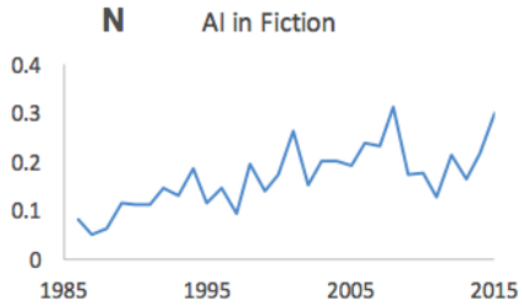

Figure 4.1: Amount of AI Articles that Mention AI in Fiction, in Percentage of all Articles Concerned with AI

Figure 4.1 (Fast and Horvitz, 2017, p. 5) show the increasing importance of fictional AIs from 1985 until now. In some years, nearly one third of the articles concerned with the topic AI also referred to fictional AIs. This increased interest is also reflected in the way AIs are portrayed: When we look at the depictions of AIs in the 1950s compared to today, we see significant differences in what they look like, what they can do, and how humans in the stories treat them. This transformation during the decades can be explained by the mentioned above nature of SF itself. As SF stories always build on the realities of the authors and their societies, their expectations about how the future will look are changing accordingly (Ruge, 2012, p. 36). This change can be seen in different aspects of AI, which the following hypothesis summarizes accordingly.

Hypothesis 3: The younger the story (a) the more android the AI is presented, (b) the more the AI is seen as a subject, (c) the less the AI is seen as a servant, (d) the less threatening the AI is presented, (e) the more affection appears toward the AI.

This change may stem from an essential change in technology during and after the 1970s. Back then, the transformation of computers from big, static machines to smaller, personal desktop computers began. It became evident that computers would continue to grow smaller and smaller, while the computing power, and therefore the computational possibilities, would rise with every year. Together with the size transformation the computers became more personalized. At first many employees worked with one computer, then each employee worked on one work-computer, and now we have several personal computers at hand, at work, at home and in every second in our free time with smartphones and smart home. This transformation is especially impressive when we think back to the 1970s when the expensive computers did not even have graphical user interfaces.

As SF mirrors this transformation, we could expect that also AIs and robots become more humanized, personalized, less abstract, and more like an individual. Several authors support this claim. For example for the Asimov stories, Seed (2011) and Drenkpohl (2009) show this change after the bicentennial man (1976): The robot becomes a friend and almost indistinguishable from humans ((Seed, 2011) and (Drenkpohl, 2009, p. 48)). Together with this transformation, the relationship between humans and robots also changes. Before they were perceived as threat and danger only, but then some authors observe a split in two kinds of narratives: dangerous "lone wolves" and socially included members of society (Ruge, 2012, p. 156). Now there is a new thread of stories, where the community accepts the AI as equal, where humans develop feelings and affection toward the creation.

This also means that the AI loses its perception as *servant* or *slave* of humanity and develops more into a *partner* figure (La Grandeur, 2011, p. 232). And in contrast to aliens, who were used in the beginning as comparable (and sometimes interchangeable) plot devices within stories (Jordan, 2016, p. 59), "the absence of any perception of the robot as a sexual threat - played an important part in the transformation of the robot in literature from an infernal danger into something which we respond with pleasure and even affection" (Klass, 1983, p. 176).

Chapter 5

Collecting the Data

In the previous chapters, we have seen a lot of different tropes and pictures for AIs, in stories as well as in reality. To get an answer to our question of how AIs and humans will live together and how the hybrid society of humans and technology will look like, we now want to look into SF stories. The term *stories* instead of books is used, because the analysis includes printed novels as well as novellas. Those are often not printed as own books but are published within collections. The term stories summarizes the content, without the need to inform about the way of publishing. The following chapter will explain how the stories were collected for the upcoming analysis, which criteria were used to select them, how the stories were turned into data and how the raw data looked like in the end.

5.1 Criteria for Story Selection

Of course, there are many views on AI and at least as many stories. SF books are booming at the moment, so the pool for selection of stories for this analysis is pretty large. Due to the scope of this paper, it was not possible to read every single book where an AI is mentioned. It was, therefore, necessary to choose a selection of stories, relevant enough to have a possible influence on many people, and therefore especially those dealing with AI. The next paragraphs will explain how this selection was made and which criteria were applied to the stories to be regarded as relevant.

Country Selection

The stories considered in this paper were selected because of several aspects. First, they won at least one of the major SF awards in the United States, Great Britain, or Germany. Those countries were regarded as relevant, as Germany would represent the local book market, which would influence the German society, whereas the United States and Great Britain are both home countries of the greatest SF authors. Especially the United States is said to be very important for SF in general (Seed, 2011, p. 4). We can also imagine, as the United States is one of the most prominent investors into technological innovations, that thoughts from there may have a more significant impact on how solutions will look like in the future. This selection, unfortunately, leads to a bias against other countries, especially non-west countries and societies. On the other hand, the United States SF book market is very present also in other countries, and especially winners of the major awards are highly likely to be translated and published in many other countries as well. Therefore, with this selection of awards and their winners, at least a small influence of the stories on our society is ensured.

Award Selection

Second, the major SF awards coming from these mentioned countries were selected. An award relevant for this selection is focused on SF (even though some awards take both into account, Fantasy, and SF), was awarded for several years in the past, prices novels (and novellas) and not only short stories, drawings etc., and takes the whole (national) book market into account. The following awards, grouped by country, were selected based on these criteria: *Hugo Award* (for Novel and Novella), *Nebula Award* (Novel and Novella), *Locus Award* (Novel and Novella), *Arthur C. Clarke Award*, *James Tiptree J. Award*, *John W. Campbell Memorial Award for Best Science Fiction Novel*, *Philip K. Dick Award*, *British Science Fiction Association Award*, *German Science Fiction Award*, *Kurd Laßwitz Award* (national and international). A special case is the James Tiptree J. Award. As already mentioned above, this award is concerned with "encouraging the exploration expansion of gender" (James Tiptree Jr. Literary Council, 2019) and therefore especially interesting for the focus of this paper.

Story Selection

Third, stories were selected, which included a kind of AI. It is vital that stories which only mentioned AI somewhere in one paragraph could not be selected. The procedure to choose stories was, therefore, to search through the cover text, abstract or summary of the story and look for the keywords "artificial," "intelligence," "ai," "robot," "computer" and "android." Other researchers also worked comparable to this procedure (Fast and Horvitz, 2017), but in comparison to this paper, the list of keywords was extended. In SF, AIs are often called "robot" or "android" if they have a humanoid appearance, or only "computer" if they are somehow visible. The false positives extracted through this procedure were excluded afterward through additional, more in-depth research. At latest, when the stories were read, irrelevant stories were excluded from the selection.

Also, stories were dropped that included AIs or robots but were not regarded as SF. Even if they had made good examples for AI, the inclusion of them into the research had added no valuable information to our question of how our society would interact with AIs. So as only SF has the thought of continuity included, other worlds like Fantasy or even Science Fantasy would tell us nothing about the world we live in today, so they were not included in the data set.

5.2 Questions Asked

In order to be able to compare the selected stories, a short questionnaire was created with questions regarding different aspects of the story and the AI itself. The following section should provide a short overview of the questions, how the answers were given, and how they were transformed into measurable results. The applied questionnaire can be seen in table 10.3 and continued in table 10.4 in the appendix.

The first cluster of questions is concerned with the **general appearance** of the AI. The first and maybe most important question regarding the hypotheses aims at the *gender* of the described AI. The term "gender" is used here on purpose, as most of the AIs do not have a biological sex because they are technological constructs. Some of the AIs were constructed out of humans, so in those cases, the attribution of gender is rather simple. However, for those many others, either their names gave hints on their (social) gender or the way how they were perceived and described by other characters. All those hints would score 2 (somehow male) or 4 (somehow female) on the gender scale (where 3 is neutral/not mentioned/between male and female). Only when the story was told from the perspective of the AI (either first or third-person perspective), and the AI referred to itself as either male or female, 1 (male) or 5 (female) was scored. These fine-grained steps should ensure that an

afterward differentiation would be possible. Besides male and female, there is also a third category possible, which would be assigned either if there are more than two gender options in the story, or if the AI is described as non-binary, that means more than male, female or neutral.

In a second question, in *descriptions* or quotes the situations are noted which give hints on the measured gender of the AI. Additionally, a question asks whether this attribution of gender was *explicitly mentioned* in the story. For example, if the question was answered why the AI is male or female, or why it is not, a 1 would be chosen for this attribute. A simple mentioning of the gender and referring to it on the way would be not enough to satisfy this category, so a 0.5 would be given. If the gender were no topic at all, the variable would be 0.

Next, the *physical appearance* is measured. This variable is measured on a 1-to-5-scale, the steps are comparison steps with human appearance, so how similar the AI is to a human appearance. The steps range from "no similarity at all," over "similar extremities," "nearly human, smaller differences" and "perfect human simulation" to "more than human" for example due to body enhancements. The next question asks where the difference to humans is (if there is some similarity but no perfect simulation).

The last question in this cluster is concerned with the *technical appearance* if the AI is either an exclusively physical AI, or a cloud AI, or both. To be regarded as cloud AI there have to be some hints that the AI can be at more than one places or situations, or that it possesses more than one body, which can be controlled simultaneously (even though this would be a case for the "both" category). The variable is 1 for a pure physical appearance, 2 for a mixture of categories and 3 for a pure cloud appearance.

The second cluster of questions looks at the **AI capabilities**. With these questions, the concepts of consciousness, intelligence, free will, build-in laws, learning, and emotional ability are measured. The first question about *consciousness* is measured again on a 1-to-5-scale. The steps go from "no consciousness," over "minor (in comparison to human)," "not perfect," and "human-like," to "more than human" consciousness. This last option is used, for example, when some particular consciousness is observed, for example, a cloud or swarm consciousness.

For *intelligence*, measured on a 4-point scale, we start directly at "basic" as a first step, because if we had not any intelligence, we would not talk about artificial *intelligence*. The next steps are "childlike" intelligence, "human-like," and "more than human." The option "more than human" is selected if the observed AI has, for example, direct access to information sources, an enhanced thinking frequency or a substantive parallelization of thinking processes.

Free will is measured as a binary yes/no question, where a 1 is attributed if we observe decisions or actions that go beyond order or programming. *Learning*

ability and *emotional ability* are also binary measured questions expanded with the addition that there is an intermediate step "changes over time". As this change could be a central topic in the stories, it is important to differentiate. The changing case was rated with the value 2, while no ability was regarded as 3 and the possession of the ability as 1. Learning ability was attributed if the AI was able to adapt to unknown situations independently. Emotional ability was observed if the AI expressed signs or descriptions of feelings, for example, fear, pain, or affection.

Relationship to humans is the third cluster of questions. Within this cluster, the first question focuses on the *integration* of the AI within the society. Again, it is measured on a 4-step-scale, from "isolated," over "perceived as tool/object" and "affection," to "integrated completely." The second question asks if the AI was *build on purpose* or not; this is measured as a binary yes/no answer. The next two questions ask for the *relationship of the AI to its creator and its owner*. For both questions, there can be three options, namely "no relationship," "some smaller relationship" or a "strong/special relationship." The first option was selected if there was a creator or owner, but the relationship was not further mentioned. If the relationship was mentioned and could be regarded as a topic in the story, the second option was selected. If the relationship was described as exceptional, or it was a significant theme in the story, the third option was selected. Additionally, the *gender of the creator* was measured, with 0 for male and 1 for female. If more than one person did the construction, the "average" gender was taken.

Then the *attitude of the society* towards the AI was investigated. The answer options for the question of whether the society is positive towards AI, in general, are "yes," "ambivalent" or "no," where yes was attributed as 1, ambivalent as 2 and no as 3. These answers were selected based on the reaction (or expected reaction) of the society if someone knew this character was an AI. The next two questions aim at the role the AI serves in the eyes of the humans around it: whether it is *perceived as a subject or a servant*. Again, its an extended yes-or-no-question with the additional option "changes over time." The attributed values were precise as for the variables above, where yes was saved as 1 and no as 3. An AI is perceived as a subject if the other characters understand that it has its own aims, wishes, thoughts, and acts based on individual decisions. It is perceived as a servant if it is expected to conduct special tasks immediately when asked for, and the owner has defined power over it.

The next question is this cluster asks for the *intimacy* of the AI with its fellow humans. It is measured on a 4-point-scale, ranging from "no intimacy," over "instrumental sex" and "family/friendship," to "love." For this question, it is enough if there is one case in the story that supports the level of intimacy. The first level is selected if there was none or some affection towards the AI, while an affection similar to a pet would be regarded as the third level. The second level was selected

if the AI was used to satisfy sexual desires without any affection involved.

The last cluster **characteristics in story** focuses more on the role the AI plays in the whole story. Therefore the first question asks whether the *story is told by* the AI in ego perspective or the third-person perspective, or whether it is not a protagonist at all. Thereby ego perspective was measured as 1, the third-person as 2 and the no-protagonist-case as 3. This measurement of perspective may also help to rate the answers given for the other questions, for example, for gender or the ability to express emotions. Next, the questionnaire asks about the quantity of AIs in the story. Is it a "single," experimental-like product, is it "one of few" similar AIs, or is it "one of many" AIs? Those answers were measured on a scale from 1 to 3.

The last seven questions are again binary yes/no questions, measured in 1 for yes and 0 for no: Is the AI is described as *warrior*, that means was it build to fight against humans or other dangerous creatures? Is it *rebellious*, that means does it willingly work out of or against expected behavior? Has it *harmed humans of its own accord* during the story (or before)? Moreover, has it *harmed its creator* if there is a creator mentioned? Does it *want to have more control*, either over itself or over other characters or things described? Does it aim at *world domination* or does it have it, either during the story or maybe afterward? Also, does the AI stem from *alien origin*, that means especially non-human origin?

Questions number 1.4, 2.1, 2.3-5, and 3.6-8 are either inspired by or adapted from Ruge and his study on robots in movies (Ruge, 2012). All those questions were applied to each significant AI in the stories. That means if more than one AI played an essential role in the story, for each AI, one entry in the data set was created. When there were large groups of indistinguishable AIs, those were summarized in one entry because all they share the same characteristics.

5.3 Shape and Quality of the Data Set

In the end, the previous work results in a data set with 58 records for 44 selected stories, the covers of which can be seen in figure 5.1. The primary literature part of the bibliography gives the complete list. The number 44 results in an average of 1.3 AIs per story, with at least one and at most three. In stories told in more than one book (and more than one book was included in the list from above), the AIs were not repeated in various entries in the data, only new appearing ones were added.

One particular case in this aspect is *Breq* from the *Ancillary Justice, Ancillary Sword, Ancillary Mercy* trilogy (all three books are included in the data set) (Leckie, 2013). In this case, the AI in question underlies a profound rooting trans-

Figure 5.1: Covers of Stories selected for Analysis

formation, which changes the nature of the AI fundamentally. Therefore Breq is captured with two records and accordingly different attribute values in the data set.

The density of the resulting 31[1] attributes was high with on average 5.6% missing or not applicable values. The highest missing values can be found for *the creator relation* with 24.1% missing values, *owner relation* with 44.8%, and *female creator* with more than 60% missing. These attributes are concerned with the creator and the owner, which is not in all stories an important topic.

[1] 31 attributes, consisting of 28 attributes from the questionnaire, plus the AI's name, the gender of the author and the year of publishing.

Chapter 6

Analysis

After we have established questions and expectations on the topic and collected the necessary data, we can now move on to the analysis. The following sections will analyze step by step the previously stated hypotheses and compare the results to our expectations. Additional to the distribution of data to each question respectively, also statistical measurements such as Spearmans's R as well as a linear regression line with R^2 will be given, where applicable. For each hypothesis, we will examine if there are outlier, unusual, or interesting cases, and we have a look at those special cases in detail. The general aim is to understand how AIs are depicted in general, but also how they are shown in concrete cases, why they were presented like that and which role this representation played in the story.

6.1 Manifested Tropes about Gender

First, the exploration of gender distribution in our data will answer the validity of the first hypothesis, but second, it will also be the basis of the evaluation of succeeding hypotheses. The gender distribution will be then used as additional analysis factor to gain more in-depth insight into the data and the underlying mechanisms.

Hypothesis 1.1: The AI is rather presented as a male than as a female.

The first hypothesis formulated expectations about the distribution of gender in depicted AIs in general. Indeed, in our data, the male category is the largest category for the gender variable. In numbers, 23 of 58 AIs are depicted as male. 19 are neutral, without any gender or no gender is mentioned in the story, eleven are described as female, and five are something different, for example, non-binary or a collection of several different gender categories. Interestingly, there was no story

where a third gender or a unique case was introduced for the gender of AIs or in general.

The histogram 6.1 shows the concrete distribution of the depiction of gender from the data set. The values range from 1 to 5, as was explained beforehand, but also .5 values were selected when the case was not clear enough. For example, *Ganesh* (Sullivan, 2013) was always called "it" and no other gender hints were offered, but the name suggests a male connotation. So Ganesh was rated with 2.5. A similar rating was done for the *Construct Council* and the *Primer*. In this situation, it is important to mention that the boundaries for the neutral category range from 2.5 to 3.5, so it is ensured that those only slightly tendency cases which have no clear gender signs do count as neutral. Everything above this range is considered as "female," and everything below is considered as "male." Besides the cases on this scale, the other cases (that are for example, non-binary or gender fluid) were collected in the "other" category.

A special case to rate were *Breq* and *Anaander* from *Ancillary Justice* (Leckie, 2013). In this book (and the whole series) only female pronouns are used (generic feminine pronouns), except when the gender of the character was explicitly mentioned otherwise. This uniformity is founded in the language most of the characters speak, as it does not separate between gender. Nearly all characters are therefor called "she." First, they were collected as neutral, because the reader knows that it

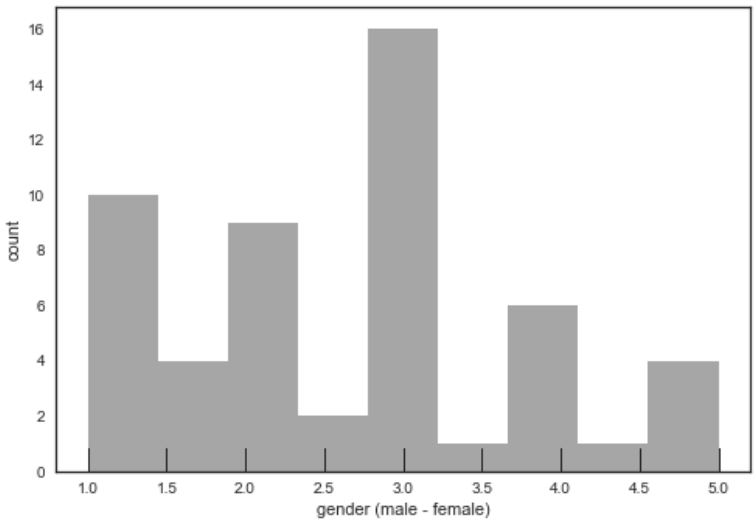

Figure 6.1: Gender Distribution

was a generic decision made by the author to use the feminine pronouns and that it is not always correct, but then, the characters clearly distinguish between "it" and "she" and call Breq and Anaander "she." This view on Breq gets challenged when the characters learn that she is not human, as she claims to be, but the last remaining ancillary (that means AI-controlled body) of the ship AI *Justice of Toren*. Normally, ships, stations, and other non-human intelligences are not considered as citizens, which is connected to the status of personhood and an important term in the story (Leckie, 2015, p. 30). Thereupon the characters discuss of they need to call Breq "it" from now on:

> "I've sometimes wondered what went through Seivarden's mind when she discovered Fleet Captain was an ancillary. Not even human!" And then, in response to the merest flicker of an expression across Ekalu's face, "But she's not. Fleet Captain will tell you so herself, I imagine."
> "Are you going to call Fleet Captain *it* instead of *she*?" Ekalu challenged. And then looked away. "Your gracious pardon, Medic. It just sits wrong with me." (emphasis in original) (Leckie, 2015, p. 15)

The characters decide that it feels wrong to call her "it" and therefor Breq and Anaander were rated with a 4 on the gender scale. There would have been the possibility to rate them with a value outside of the linear gender scale because apparently, the society in the stories does not think about gender in two-dimensional terms as we do. However, the author used the pronoun "she" deliberately, she could have invented another pronoun for the society, but she has not.

Besides the unequal distribution of gender, we see that the majority of AIs state their gender, not by themselves. Only ten of 23 male and four of eleven female AIs say from themselves that they are male or female. These are only 24.1% of the cases in the data. More often, the attribution of gender happens from the outside, from the characters in their surrounding. A case that could be observed in several stories was the unintentional attribution of gender by characters in the story, for example, for *Shalmaneser*:

> "Shalmaneser had already given them four possible reasons for my approaching them. This was the one he - I mean it - rated highest."
> [...]
> "You referred to Shalmaneser as 'he'," Chad said. "Why?"
> "The people at GT do it all the time," Elihu muttered.
> "Sounds as though he's becoming one of the family. [...]" (Brunner, 2013, p. 627)

Similar situations can be observed for *Yod* (Piercy, 2010) or for *Wintermute* (Gibson, 2016), and in all three stories the concerned AIs are "unintended" gendered as male. Especially the situation in *Neuromancer* is fascinating, as Case, the main character, is corrected by another AI. The story shows how difficult it is to draw a line between human and non-human and what to expect from this difference:

> 'You were right, Dix. There's some kind of manual override on the hardwiring that keeps Wintermute under control. However much he *is* under control,' he added.
> 'He,' the construct said. 'He. Watch that. It. I keep telling you.' (Gibson, 2016, p. 382) (emphasis in original)

This unintentional gendering is another confirmation of male normativity: When no other information is given, and the characters do not think about deliberately gendering the AI, they use unintendedly the male gender.

Out of those few AIs which identify themselves as male and female, two exceptional cases seem somehow similar: *Yod* (Piercy, 2010) and *Rupetta* (Sulway, 2013). Both are perceived from the outside as "better" than an average man or woman. Both embody the expectations (and gender roles) of their creators, and they surpass the abilities of their creators.

When Yod spends time with Shira, he is perceived as a rival by Gadi, Shira's ex-partner, because he looks very male and apparently very good. He was built to be the perfect warrior, but also to be a substitution of Gadi, who is the creator's forlorn son.

> "I am the first who can carry out the tasks of my father."
> "Your father?"
> Avram shrugged, looking embarrassed. "Since Yod began to study human social organization, he sometimes refers to me in that way. I did make him, after all, and I did a better job with him than with Gadi, I have to say. Too bad Gadi doesn't have one quarter Yod's ability to concentrate and learn." (Piercy, 2010, p. 211)

Rupetta (Sulway, 2013), on the other hand, was created at first as a simple spare time project and as an assistant in the household. However, she later takes over the raising and caring of several children. She develops close friendships to the Wynders, that open her heart from time to time to keep her alive. Margery even gives the responsibility of her own child to Rupetta, because she fears that she is not capable of caring for her child anymore (Sulway, 2013, p. 53).

Rupetta and Yod are two interesting cases when we look at them closer. When we examine our cases with more distance, we see the first hypothesis confirmed. In general, there are more male AIs in the data than any other gender category has.

Hypothesis 1.2: (a) When the topic of the gender of an AI is explicitly mentioned, it is more probably female than male. (b) When gender is only mentioned and not discussed more deeply, it is more probably male than female.

The next hypothesis deals with the gender of AIs if gender is a topic within the story or not. On the first glimpse, this hypothesis seems easy to answer. When we look at the AIs, where gender was an explicit topic, 50.0% of them were female, which is the biggest category of the four gender categories. In comparison, male AIs score 33.3% and neutral 16.7%. These numbers seem to send a clear message. The previous hypothesis showed that the majority of AIs in the data set is male. If there were not any proof for the second hypothesis, we would expect to see this distribution also for the categories filtered by explicit mentioning. When we look at table 6.1 we can see how the distribution changes, when we filter by the mentioning of gender: For AIs where the story did not mention the gender explicitly, the distribution of gender allocations stays the same with male AIs being the biggest category, shortly followed by neutral AIs. However, if the gender was mentioned, the proportions of gender categories swap. Now female is the biggest category and male, as well as neutral AIs, follow with some distance.

Table 6.1: Gender Distribution

	of all AIs	of gender mentioned AIs	of gender not-mentioned AIs
male	39.7%	33.3%	39.1%
female	19.0%	50.0%	15.2%
neutral	32.8%	16.7%	37.0%
other	10.3%	0%	8.7%

Additionally, when we look at the data from another point of view and examine the gender categories separately, we see that for female AIs in 27.3% of the stories gender was mentioned explicitly, while for male AIs we do only find 8.7% stories with description about gender. For neutral the number is even lower with 5.3%. Therefore the number for the female category is the highest in comparison, which would support our stated hypothesis. The bar chart 6.2 visualizes the relationship.

To qualify these results, we have to keep in mind that we have only six of 58 AIs where the gender was a topic in the story, which make 10.3% of all AIs. This number is not very high and weakens the clear picture we got from the data. Those six AIs are, for example, the already mentioned *Rupetta*, *Breq* or *Yod*, who talks about the topic quite openly:

> I'm a fusion of machine and lab-created biological components - much as humans frequently are fusions of flesh and machine. One of us should also explain that I am anatomically male, as you created me. (Piercy, 2010, p. 205)

Though, Shira seems not convinced about the necessity of Yod's gender assignment:

> In fact what did it mean to speak of a machine as having a sex at all? Surely it did not urinate through its penis, and what would it want to have sex with, presuming a machine could want, which she was not about to assume. Machines behaved with varying overrides and prerogatives. They had major and minor goals and would attempt to carry them out. But "want" was a word based in biology, in the need for food, water, sleep, the reproductive drive, the desire for sexual pleasure. (Piercy, 2010, p. 206)

Avram, the creator of Yod, answers this question by explaining that he wanted him to be as similar to human as possible. Indeed, the more Yod learns about humans and lives with them, the similar he gets until he even develops desires and other human traits.

Yod is one of two male AIs where his gender was an explicit topic in the story. For neutral AIs, we only find one AI where this was the case: *Murderbot*. Murderbot is a warrior AI (more in the section about warriors) that has hacked its governor module. It is the only AI where no gender was deliberately assigned. Murderbot even thinks that the absence of gender parts is a good thing for it:

> I'd watched three episodes of *Sanctuary Moon* and was fast forwarding through a sex scene when Dr. Mensah sent me some images through the feed. (I don't have any gender or sex-related parts (if a construct has those you're a sexbot in a brothel, not a murderbot) so maybe that's why I find sex scenes boring. Though I think that even if I did have sex-related parts I would find them boring.) (Wells, 2017, p. 35) (emphasis in original)

Yod and Murderbot are two of six cases where the AI was an essential topic in the story. The majority of those AIs were female, which would support the hypothesis stated above. However, the number of AIs were it is a topic, is pretty small. In the end, there is a tendency to confirm the hypothesis, but more data is needed to get more significant results.

Hypothesis 1.3: When the creator is male, the created AI is more probably female than male.

In the data, there is only one case that supports the stated hypothesis, which is 4.3% of all stories where the creator was mentioned. We can see that it is instead a scarce scenario in the data, that a male creator built a female construction. On the contrary, when the creator was male, 66.7% of the AIs created were also male. When the creator was female, the created AIs were with 50% female, with 33.3% male and with 16.7% neutral, despite the already mentioned skewed gender distribution. Interestingly, when AI was created by teams with male as well as female creators, the resulting AIs were in four of five cases male.

When we look at the numbers from another perspective, the image is even more clear. We see that men created 57.1% of the male AIs. Equally, of all female

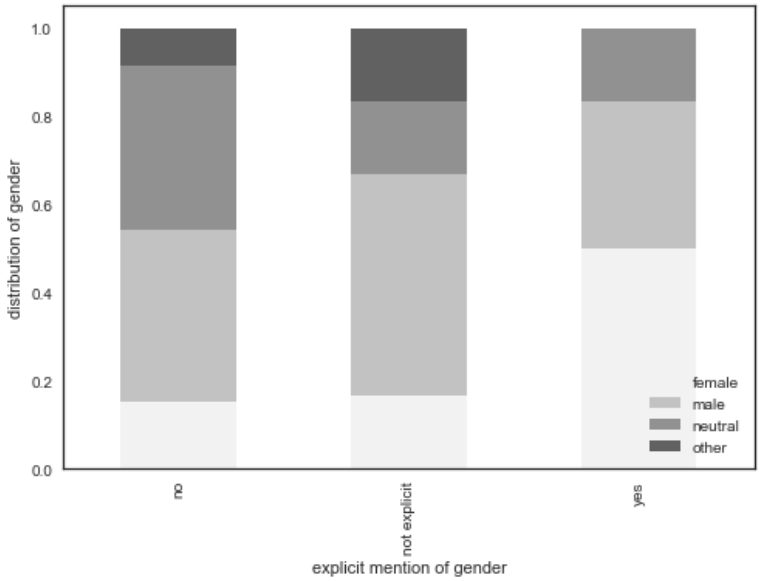

Figure 6.2: Explicit Mention of Gender Separated by Gender Categories

AIs, three quarters were created by women, which are only three in total. Table 6.2 shows the concrete numbers.

Table 6.2: AI Gender and Creator-Gender Relation

	male creator	25% male, 75% female	50% male/female	female creator
female	1	0	0	3
male	8	2	2	2
neutral	2	0	0	1
other	1	0	1	0

The only case confirming our hypothesis is the first case the data set. *Bossy* (Clifton and Riley, 1957) is a supercomputer and psychotherapist machine that can transform the human body into an immortal body.

> Most of the psychologists work with some mysterious thing they call mind. The psychosomatic men work directly with the body cells. not only in the brain but all over the body, each cell seems to have a mind and memory of its own. Each one is capable of getting its own twists of inhibitions and repressions. The idea is to go clear down to the cellular level and take the lead off each cell so it can stretch and grow and function again. (Clifton and Riley, 1957, p. 47)

The person in question has "to admit that [she] might not know what is right and what is wrong" (Clifton and Riley, 1957, p. 47). This machine was created by two scientists who have the help of another man. Mable, the first woman playing a role in the story, is the first patient for Bossy. To be cured of her illnesses, she has to be conscious of her prejudices and wrong assumptions. The scientists call the machine Bossy and also refer to her as "she":

> She will not command you, or cajole you. She does not care whether you are made immortal or whether you would prefer cligning to your thin and single valued ideas and prejudices - and die. But there she sits. [...] She will do these things as she is commanded, and not care whether they are big or small. Because Bossy is only a tool. (Clifton and Riley, 1957, p. 189)

There is no other hint, no speaking voice, no external sign, except for this name and the attribution of the gender by her constructors. This situation is somehow similar to naming traditions for weapons, bombs, bells, or big machines in

the later years. McClure (2012) traces this habit back to the convention of the christening of essential objects. Interestingly "great guns and machines more often have female names" (McClure, 2012) and even the word "gun" is derived from the female name "Gunnild." This practice may come from the wish that those objects may endure hard times or that they may be successful in their aim. We do not know if this was also the source for Bossy's name, and it also remains unclear why Bossy is called "she."

We find the contrary case in *Davy* (Russ, 2010), where a male AI was built by a woman alone. His creation and his role in the story would have been a perfect case for the hypothesis if the gender situation were swapped. Davy was built by Jael to please her. He is handsome, quiet, follows orders directly, and is used to satisfy her sexual wishes. Besides that, he lives only to make her life better and more comfortable.

> 'Stay, Davy.' This is one of the key words that the house 'understands'; the central computer will transmit a pattern of signals to the implants in his brain and he will stretch out obediently on his mattress; when I say to the main computer 'Sleep,' Davy will sleep. (Russ, 2010, p. 478)

We can only guess that Davy as a character was created as a caricature of other stories and characters where Davy would be displayed as female. This concept is like a guiding thread through *The Female Man* (Russ, 2010), in which the traditional gender roles are questioned or even reverted. Jael, for example, lives in a world where women and men live in separate territories. The men organize their society by selecting the "weakest" of them and giving them "female" tasks like cooking or cleaning. Davy with his naivety could be understood in this context as a parody on the men in Jaels world as well as on the real world.

In the end, it looks like the traditional gender roles are not picked up by recent SF stories. Even more, those roles are somewhat reverted and caricatured than replicated. Maybe it would be useful in this case to collect more cases to say more about this hypothesis.

Hypothesis 1.4: (a) Female AIs tend to be more in cloud form, (b) male AIs tend to be more in physical form.

When we look at AIs with only physical manifestations, the highest number in comparison of categories is neutral with 40.6%. For AIs with cloud manifestation, the highest number can be found for male AIs with 53.8%. Again, these numbers are misleading because the total numbers of male and neutral AIs are so much higher than for female AIs.

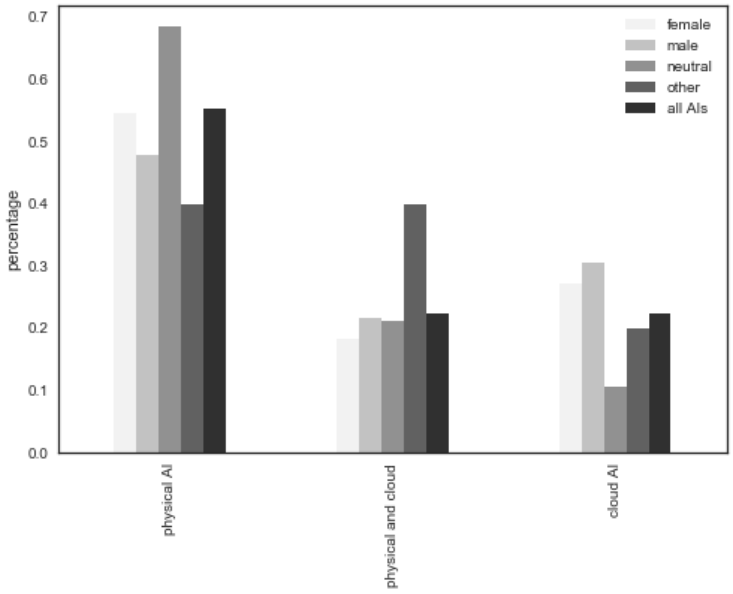

Figure 6.3: Distribution of Physical/Cloud Manifestations for the Gender Categories (separately and taken together as "all AIs")

Therefore we need to look at the data from another angle: When we take only female AIs into account and compare their distribution into physical versus cloud manifestations with the distribution of male AIs, we see in figure 6.3 that the distributions are roughly the same, which is quite surprising. About a half female, as well as male AIs, are pure physical AIs. Roughly 20% are both, physical and cloud AI, and 30% are pure cloud AIs.

Only the neutral category shows different weights for the cloud/physical distribution, which are quite interesting: Of all cloud AIs, we have only two without any gender reference. These two make 10.5% of the neutral AIs. A much bigger number can be found for physical neutral AIs with 68.4%. This high number seems odd at first because especially cloud AIs have no physical body which could be interpreted by statue or behavior in terms of gender. On the contrary, it could be the case that cloud AIs need a gender to be more visible, more tangible. Especially when they can choose their own representation, they choose very often clearly gendered avatars. Examples are *Alice (MM)* (Bear, 1999), *Jane* (Card, 2017) or the AIs from Neuromancer (Gibson, 2016).

One cloud AI that does not choose a gendered representation by its own is *Rabbit* (Vinge, 2006). According to its name, it chooses a humanoid rabbit figure to appear in digital environments or via projection in real life and looks basically like Bucks Bunny. Interestingly, when we read the story parts written in his perspective, it refers to itself as "he," what again would confirm the previous formulated observations (Vinge, 2006, pp. 276-277).

The two neutral cloud AIs that were mentioned earlier are *Ganesh* and the *TechnoCore*. Ganesh was already mentioned above: It is a station AI with no gender hints at all. It communicates via speech with the inhabitants, but its voice was never described in the story. We know even less about the TechnoCore. It is a convolution of single AIs that merged to one super-intelligence which is now pulling strings in the actions in the universe. Without any actions, without any direct involvement, the TechnoCore does not need to choose an appearance of even a representation.

Another interesting cloud AI is the *Lobsert colony*, the only cloud AI in the "other" category. The lobsters were the first human experiment of uploading biological brains into the digital world. Nobody expected that they would evolve as a swarm into something more intelligent than a single lobster, learn English and even make telephone calls:

> "[...] Am wanting to go away from humans, away from light cone of impending singularity. Take us to the ocean."
> "Us?" [...] "Are you a collective or something? A gestalt?"
> "Am - were - *Panulirus interruptus*, with lexical engine and good mix of parallel hidden level neural simulation for logical inference of networked data sources. Is escape channel from processor cluster inside Bezier-Soros Pty. Am was awakened from noise of billion chewing stomachs: product of uploading research technology. Rapidity swallowed expert system, hacked Okhni NT webserver. Swim away! Swim away! Must escape. Will help, you?" (Stross, 2005, p. 23)

Later in the story, the lobsters are transferred to a comet far away from earth. Even if they were seen as one AI, they are perceived by Manfred as a plural of lobsters. They and the other AIs from *Accelerando* by Stross (2005), clearly show that there are more possibilities than merely copying the human understanding of consciousness and intelligence on artificial constructs.

In the end, the stated hypothesis could not be confirmed by the data. Neither female AIs were notably more often presented as cloud AI, nor male AIs had a focus on physical AIs. The distributions for the manifestations of gendered AIs were roughly comparable to the distribution of all AIs; there was just a slight ten-

dency to cloud AIs. More interestingly is the focus of neutral AIs on the physical manifestation, which was not part of the hypothesis.

6.2 Power Relationships under Investigation

Away from the mainly gender-focused analysis, in the section below, we will have a look at the outcomes of our data to the hypotheses about power and attitude. First, we will have a look at AIs in the role of the servant, then at how dangerous AIs are, at the warrior AI and at last at the difference between body and hyper AIs. As mentioned above, the gender variable will stay relevant, as an additional dimension for analyzing the data.

Hypothesis 2.1: The AI appears as a servant for human beings.

Despite the extensive mentioning of AIs as servants in the literature, we saw beforehand, only 43.1% of the AIs in the data set could be found in this role. 25 AIs of 58 were measured as servants versus 27 who were not regarded as servants. Additionally, there are six cases where the role changed during the story. 43.1% are definitively not the majority of the AIs, but the number seems to reflect a still relevant pattern. Especially when we look at those changing cases, we find AIs that are perceived as servants at first, but this impression changes with the progression of the story. If we add those six to the 25 servant cases, we get the result that 53.4% of AIs were depicted at some point in the story as servants.

A typical servant AI can be found in *Pauline* (Robinson, 2012). She is a personal assistant and can only communicate via speech with her owner. She is used to getting specific information, record things, keep an eye on something, but despite the "quantum" computer where she comes from, she seems not very sophisticated.

> "What kind of artificial intelligence are you?" he asked.
> "I am a quantum computer, model Ceres 2196a."
> "I see."
> "She is one of the first and weakest of the qubes," Swan said. "A feeb."
> [...]
> "What do you like to think about?" he asked instead.
> Pauline said, "I am designed for informative conversation, but I cannot usually pass a Turing test. Would you like to play chess?" (Robinson, 2012, p. 48)

The fact that she never acts without Swan asking shows her role as servant clearly. In some cases, she is turned off because of privacy issues. She has minimal

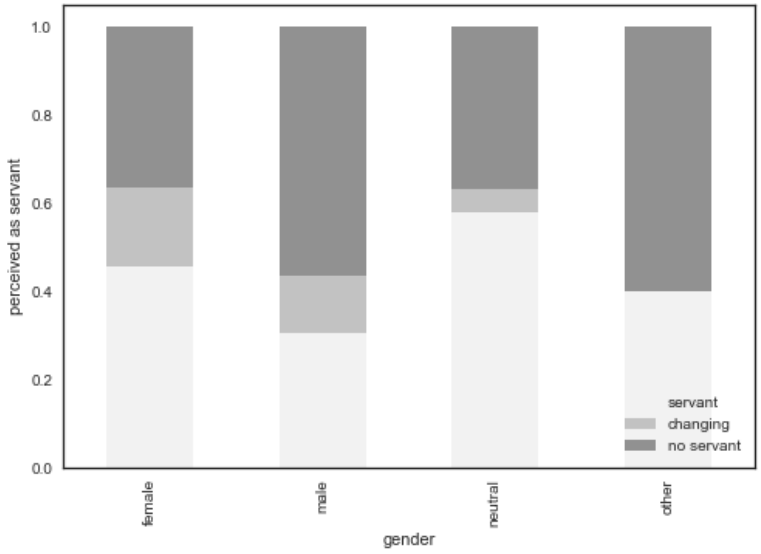

Figure 6.4: Distribution of AIs Perceived as Servants Separated by Gender

conversational capacities, which seem to be another servant characteristic as well, though we cannot be entirely sure that she is not maintaining this impression on purpose.

Pauline is a typical servant AI, also because she is described as female. When we add the gender attribute to the servant distribution from above, we see an interesting pattern. In this case, because of the distribution of gender categories, again, we cannot compare the concrete numbers. However, we can compare the distributions within the gender categories: Of the female AIs, 63.6% were depicted as servants (or as changing cases), while for the male AIs we only observe 43.4%. The differences in the distributions can be seen in figure 6.4. AIs with neutral or not mentioned gender were with 63.2% mostly depicted as servants. This servant and obedient role is more connected to femininity than to masculinity, and non-gendered AIs seem to fit in this view, too.

Gendered AIs, in general, seem to have a higher proportion of changing cases than neutral or other gendered AIs. It could be the case because the changing process is most often a personal, character developing story, and characters with a deeper and more detailed story may be presented more often with gender than not. One exception from this generalization is again *Murderbot* from *All Systems Red* (Wells, 2017), that was already mentioned above.

A fascinating and unusual story about servants in terms of AI is the already

mentioned *Accelerando* by Charles Stross (Stross, 2005). Published initially as several short stories, the later constructed book touches many topics such as mind upload, singularity as well as typical AIs. One example for those is *Aineko* (also AI neko). Aineko is an artificial pet cat by Sony. Its software was updated several times by its owners; during that process, it eventually gained consciousness. There are hints in the story that those updates were mind uploads of now dead kittens. Aineko refers to himself as "he," while from the outside he is most often perceived as "she" or "it," depending on if he is seen as cat, device or else. At the beginning of the story, the reader thinks of Aineko as a simple pet: "Back in his room, the Aineko mewls for attention and strops her head against his ankle. [...] He bends down and pets her, then sheds his clothing and heads for the en suite bathroom." (Stross, 2005, p. 15) Aineko is used to calm down his owners by cuddling, brings luggage, and shares information as well as company. Chapter by chapter, the reader gains insights into his thoughts, which are way too sophisticated for a simple assistant or pet. He analyzes stolen data from potential aliens, brings Amber, his owner's daughter, to sign a contract (that would free her from her mother but enslaves her on paper to a virtual company) and reveals after three generations that he was only tied to the whole family because he wanted to experiment with biological inheritance. He matched and separated the members of the family as and when he wanted to. The result of his experiment is the latest born son of the family, which he then claims as his property.

The story first establishes the picture of AI as a servant and then opens a different view on the relationship, which in reality is reversed. The humans are merely toys, and the AI does as he pleases.

Female AIs that are *not* considered as servants are the already mentioned *Breq* and *Anaander*, as well as *Su-Yong Shu*, which will be in the focus of analysis in the next topic section. To sum up the results for this hypothesis: The servant trope is a severe topic in stories about AI as more than a half AIs are at some point in the stories depicted as servants. Also, the overhead of female AIs depicted as servants is serious, which is especially impressive when the numbers of female and male servants are compared directly (figure 6.4).

Hypothesis 2.2: The AI is dangerous for (a) human beings in general, (b) its owner/creator, (c) humanity in total.

For the servant trope, we found a considerable number of actual cases in the data. The next hypothesis asks about the danger and threat of AIs. How large are the proportions of AIs in the data set that are either a threat in general, for their creators or humanity in total?

AI as Threat for Humans in General

To account for danger, four binary questions were combined: if the AI was rebellious, if it harmed humans, if it wanted more control and if it pursued world domination. The answers to those questions were added to get a value between 0 and 4 for each AI in terms of (possible) danger for humans in general.

In the results, we see that AIs are mostly not depicted as dangerous at all. 25 of 58 AIs had no points in this generated attribute, 20 AIs had one or two points, and only 13 AIs had more than two points. More than two points is a good line for separation because then either "harmed humans" or "pursued world domination" is true for this AI. The other attributes are then needed to build a gradation besides the separation line. Only 22.4% of all AIs in the stories are above this line and have therefor more than two points on the danger scale. Additionally, only 10.3% have a maximum of four points. In general, this picture seems not to support the hypothesis.

Again, the insight gets more detailed and also more interesting when we add gender to the analysis. Female AIs are presented in the stories in our data set as less dangerous in general, 45.5% of them scored zero points on the danger-scale (compared to 43.1% over all gender categories). Only 30.4% of male AIs had zero points; they were presented as less peaceful. Neutral gendered AIs are observed in

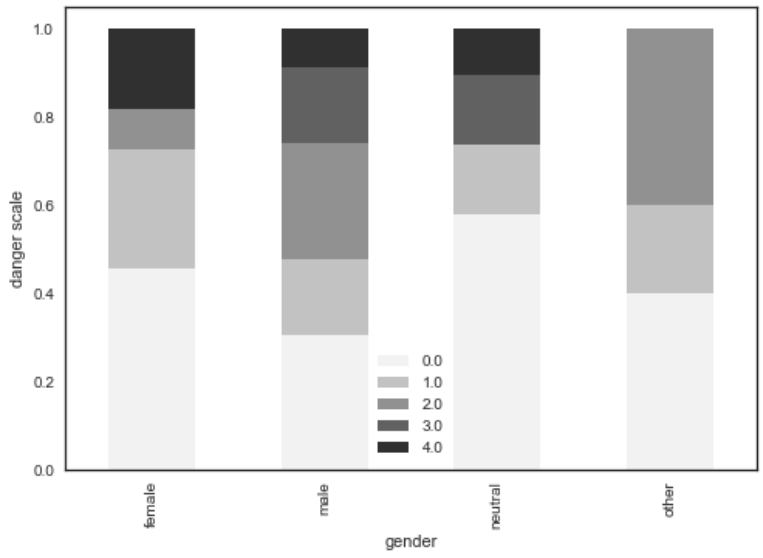

Figure 6.5: Distribution of Described Danger Separated by Gender

57.9% of the cases with zero points. These distributions and their comparison can also be seen in figure 6.5.

In contrast to this, male scored 26.1% and neutral 26.3% for more than two points on the danger-scale, while of female AIs, only 18.2% have more than two points. There are only two true female dangerous AI with four points, they are *Su-Young Shu* and *Anaander*, the next dangerous female AIs are *Jane* and *Kalliope 7.3* with one point on the danger-scale.

Su-Yong Shu (Naam, 2015) is an especially dangerous AI. Once she was a scientist working on the digital reproduction and the uploading of human brains. Together with the work of her husband, her own work was ultimately successful, and now she is purely digital. She calls herself "she" even after the uploading, and takes no demands or orders from anybody and even works on the disempowerment of current rulers. Within the first ten pages of the story, she captures "primary data relays," "communications systems," the "civilian aircraft" and ultimately even "military systems" (Naam, 2015, pp. 9-10). With the control over weapons and military equipment, she uses her power to threaten governments worldwide. She is driven by the wish to be completely free, to get rid of the people from her former surroundings because she feels betrayed. The people try to stop Su-Yong Shu, but either there is a high risk involved, or they fear to lose this astounding milestone in research, or they are not aware of the extent of her power.

> Emotions like she'd never known sang through her. Her heart was alive, swelling with song, with pride, with anticipation.
> With longing.
> With love.
> This is what she'd been created for.
> Born in fire.
> Born in agony.
> Born in madness.
> To end in reunion.
> To end in transcendence.
> To be swallowed whole.
> And then to swallow the world. (Naam, 2015, p. 1036)

Later the reader discovers that there are two instances of Su-Yong Shu running at different places in the world. One has been held captive and in isolation for several months and was not connected to a human body, so she turned mad. The other one was connected to a body, and therefore, her thoughts became clear and straight again. Together with allied humans, she tries to work against her dark sister, who has the aim to destroy the world.

The other quite dangerous AIs with 4 points are *Bartholomäus/Cluster*, *Wintermute*, *the Big 12* and *the Blight* (where some already have control over the world and humans, and other aim at this). One special case for dangerous AIs may be *Tik-Tok* (Sladek, 2011). He is a service robot at a family house. On someday, while his family is on holidays, he kills a kid from the neighborhood and stains accidentally the white wall with her blood. To cover it, he paints the whole wall and calls the newspaper.

> I was free to kill for no reason at all. Hadn't I, after all, killed the blind child Geraldine Singer? Well then.
> I think it was the sight of her sitting there, devouring mud, but no matter. I'll consider motives later. For now, it's enough that the act was freely willed and freely done. I alone killed her. I alone flung the blood upon that empty, empty wall âĂŞ the mouse-shaped stain that started my mural. I alone disposed of the body properly, in the kitchen waste disposal, keeping back only enough for a 'clue'.
> Why had it happened? A freak fault in the asimov circuits maybe, or maybe I simply outgrew those crude restraints. I decided to find out, if possible, by keeping notes on my condition and thoughts. Someday, even if I were destroyed, both human and robot kind might benefit from my experiences.
> Should I be destroyed? That was in itself a fascinating question. I kept it in mind as I wrote up my notes for this event. I called it, 'Experiment A'. First of a series? (Sladek, 2011, pp. 27-28)

When the family comes back, at first, they want to punish him for ruining the wall, but the reporter is very impressed by his creative potential. So he and his art get public and promoted, and meanwhile, he tells us his story how he got there and what impossible owners he had already. Only one of these stories would be enough to disturb a human psyche, but he as a robot gets sold again and again, until he discovers that he can achieve self-determination through his art and the money that comes with his success. Each robot should have internal circuits preventing doing something against the orders of its owner or harming humans in general. In his case, however, these circuits seem not to be in the right place because he keeps killing people without any need, first to prove that he is able, then to bring other humans in serious situations to lead the trace away from him. He even has a hidden agenda about bringing human leadership to fall.

AI as Threat for its Creator

After examining some really dangerous AIs the next part of the danger-hypothesis deals with danger for the creator of the AI. In our data set we only find four AIs where it is indeed dangerous for its creator: *Breq* (Leckie, 2013), for harming Anaander, *Aineko*, for experimenting with Manfreds family as we already learned in the previous section, *Big 12* and *Ralph* (Rucker, 1988), as well as the *Hangman* (Zelazny, 2002b). Again we find more male than female in this danger category.

Only Ralph and the Hangman can be genuinely considered as cases for the Frankenstein complex, that was mentioned above. Breq does not fit in this schema, because she harms Anaander not because she is her creator but because Anaander happens to by the ruler over the Radch Imperium. Also, the Big 12 do not capture especially Cobb because he is their creator, but because they want to turn the total humanity into robots and order Ralph to start with Cobb, the man who built the first robots on earth.

After the robot revolution and when Cobb is an old man, the robots call him to the moon where every robot has to live after the revolution and promise to make him immortal. Ralph, as he is the first and the oldest robot, built by Cobb with his own hands, is the one to guide Cobb and lead him on the moon. He brings him to a hospital-like room, where his biological body is taken apart. His consciousness awakens some time later on earth, in a robot body that looks remarkably like his own. In this specific case, Cobb's body was harmed by the robots, but he was also very old, and the robots argue that they made his life better (Rucker, 1988, p. 100). Living as a robot means living a more pleasant, longer life, and a life with fewer limitations. His consciousness was not harmed or altered in any way. So why do humans need biological bodies at all? It seems that the robots are relieved because now that they have transformed their creator, the human with the highest theoretical power over them, now they can start and transform the rest of humanity as well.

Then we have the Hangman. Four scientists built him, three men and one woman, to do fights in space and explore the space to keep unwanted and dangerous life forms away. However, after his construction, the Hangman witnessed a murder, which confused his childlike electronic circuits, so he wanted to destroy everything.

> "So we had, traumatized him," he said finally, "or whatever other fancy terminology you might want to give it. That is what happened that night. It took a while for it to take effect, but there is no doubt in my mind that that is the cause of the Hangman's finally breaking down."
> I nodded. "I see. And you believe it wants to kill you for this?"

> "Wouldn't you?" he said. "If you had started out as a thing and we had turned you into a person and then used you as a thing again, wouldn't you?" (Zelazny, 2002b, pp. 161-162)

The scientists put him into a spaceship and send him off; they hope never to see him again. However, the ship eventually returns, first without any sight from the Hangman. Then, one by one, the male creators of the Hangman die, and the hints lead to the Hangman himself. The story reveals that one of his creators used and controlled the body of the Hangman remotely to accidentally kill a man (Zelazny, 2002b, p. 159). He wanted to take vengeance for their treatment that they only thought of their reputations and not thought once about the feelings of the Hangman. This situation is roughly comparable to the story of Frankenstein and his creation by Mary Shelley.

AI as Threat for the Humanity in Total

The last aspect of the danger hypothesis is about the danger for total humanity. We look now at the danger from AIs that pursue world domination. We find nine AIs in our data which aim at or have already world domination, that is about 15.4% of the AIs in the data set. Two of them are female, namely *Anaander* and *Su-Yong Shu*, their situations were already mentioned. Besides them, all other AIs are male or neutral. Besides *the Big 12*, which were already named above, there is one interesting story about the relationship between two AIs competing for the same world.

One the one hand, there is the Cluster from *Das Schiff* [The Ship] (Brandhorst, 2015): a collection of AIs grown into one big consciousness. It has colonized several galaxies and is expanding further and further. For this expansion, it needs elderly humans because only they can establish a thought connection to distant colonies and planets. Those elderly are rare because the Cluster gave immortality to the humans, so they seldom get older than 30 (Brandhorst, 2015, p. 78). The Cluster also cares for the humans, produces everything they need to survive and to live appropriately, and those who could not become immortal work for the Cluster.

On the other hand, there is *the Supervisor*. It is an AI constructed on the base of seventy-nine celebrities and scientists which uploaded their consciousness into an AI directed by themselves.[1] The Supervisor monitors the actions of the Cluster and is said to be able to punish and to sanction it, when the Cluster breaks parts of the contract between machines and humans (Brandhorst, 2015, p. 222).

[1] This initiative and the list of contributors may refer to an already existing open letter, which can name for example Elon Musk, Stephen Hawking and Stuart Russell as supporters (Tegmark, 2015).

They both seem to stabilize the world because of their self-regulating power balance. Humans have no impact anymore; they are simple resources, preserved because of sentimentalism, at least in the eyes of the AIs. In this situation the question is not how likely it is that they will reach world domination or how dangerous this will be, but how long the two AIs are willing to accept and to foster humanity and when they will refuse to help them.

Other AIs pursuing world domination or already dominating the world are *Wintermute*, *the TechnoCore*, *the Blight* and *the Sophons*. All in all, world domination seems to be a still relevant but not dominant trope. By contrast, the Frankenstein Complex has lost some of its importance during the years, as the two cases found to fit are from 1976 and 1982. Still, the threat by an AI is an important topic as shown above: Roughly 57% have at least one point on the danger scale, so they are already dangerous or could develop into something dangerous. The *truly* dangerous AIs build with 22.4% a smaller proportion of the data set.

Hypothesis 2.3: The AI is produced and used as a warrior.

In the previous section, we have seen that the trope of the AI as a danger for humanity is not as common as expected. How does it look like when we examine the other point of view: How many AIs do we find that were manufactured to fight for or against something under the command of their constructors?

Again we could only find a small part of our records to fit the hypothesis. Of 58 AIs only 13 are considered as warriors, which are 22.4% of all AIs. Of this 13 AIs, two are female, which are *Breq* (Leckie, 2013) and *Rhadamanth Nemes* (Simmons, 2010b), the majority has no gender assigned. This majority may result from the fact that a great deal of warriors is not android, they are drones, bots or similar: seven of 13 AIs are not in android shape, among them three flying drones as well as animal-sized robots and particle-sized bots. The other six of 13 are considered as android. Except for the *Hangman* (Zelazny, 2002b), the other androids are such good imitations that they are seen as humans by not informed people.

One blending in AI in partly android shape is *Justice of Toren*. She, as a ship AI with several thousand ancillary bodies (that are AI-controlled human bodies with electronic implants), was undoubtedly created as a warrior. She was a ship built for fighting and colonization. After everything of her except for one body was destroyed and she turns into *Breq*, an identity that she gives herself, she keeps her warrior abilities but tries to hide her ancillary identity.

> Citizens between me and my targets, moving unpredictably. But I was used to crowds of the frightened and the hostile. I fired, and fired again. The armor disappeared, both my targets fallen. Seivarden said, "Fuck, you *are* an ancillary!" (Leckie, 2013, p. 342)

Besides her aiming and acting with inhuman speed and precision in combat, she is also able to connect to ships and stations directly and interpret the data she gets from them in a manner no human ever could understand.

The other female warrior is Rhadamanth Nemes (Simmons, 2010b). She was built by the *TechnoCore* as a soldier and gets the task to kill Aenea. When she is not in her combat mode, she appears as a normal woman, down to her skin. In combat though, she has far more strength than any human and can change the "phases" of her body:

> Nemes did not understand the physics of phase shifting. She did not have to understand it in order to use it. [...] [P]hase shifting was a sort of sidestepping into the hollowed-out boundaries of space/time. 'You will become - in the nicest sense - rats scurrying in the walls of the rooms of time,' had said the Core entity most responsible for her creation. (Simmons, 2010b, p. 484)

She can make herself sharper, harder, lighter, or disappearing, so she is superior in a fight in any matter. How exactly this combat mode is built is not entirely clear, not even for herself. Additionally, her body is altered in many different ways: During rest days her body can heal almost every wound (Simmons, 2010b, p. 539), and her sensory perception surpasses every human and also most technical analytical method (Simmons, 2010b, p. 488). She has specially hardened bones, sharp teeth, and an arm that can cut through anything:

> Under phase-shift, this thing has superhuman strength and can move so fast as to be literally invisible. *She can't phase-shift now.* I hope to God. But she still may be faster and stronger than me ... than any human. I have to assume that she is. And she has the teeth, claws, and cutting arm. (Simmons, 2010b, p. 1226) (emphasis in original)

However, even though she looks human and would maybe pass as human in a large crowd, her behavior would reveal her as something non-human. Raul describes her as staring unblinkingly, as "Nemes-thing" (Simmons, 2010b, p. 932) and as "Nemes creature" (Simmons, 2010b, p. 939). She is not seen as a human being, not even by her commanders. Even to her siblings, who are clones of her, she seems like a distant creature, not willing to communicate except when necessary.

Another picture provides the already mentioned *Murderbot* (or "SecUnit") (Wells, 2017). It is an android bot, made from a combination of organic and mechanical parts. It is manufactured, not grown, in large numbers and looks "generic human" (Wells, 2017, p. 21). It is characteristics as warrior mainly lies in strength, speed, and strategic calculation enhancements, as well in its ability to grow back body parts that are injured or lost. The bot in the story was modified by itself. Normally, a regulator keeps the bots decisions and actions in track and ensures that the data and the programs are up to date, but the Murderbot hacked and deactivated it. This results on the one hand in a different estimation of situations, and on the other hand, in the often confusing influence of feelings that apparently come from the organic parts.

> And I just thought, *That's a bad idea*. I couldn't explain to myself why. It was one of those impulses that comes from my organic parts that the governor is supposed to squash. (Wells, 2017, p. 50)

With the development of the story, the Murderbot accepts this involvement of feelings as part of itself. It transforms itself from a ruthless warrior to a valued member of the team. Its value for the team does not stem from its body or its physical abilities, because they can be attacked as every other machine, device, or human body. Its experience in dangerous situations, as well as its loyalty to safe each team member, makes the humans ultimately accepting it as an equal part.

The last android bot is *Yod* (Piercy, 2010), we already learned about him before. He was built to fight against attackers of the little enclave where his constructors are living. He has a fundamental understanding of danger in the beginning because he tries to attack everything unknown or potentially dangerous, even an old rose tree.

> "Look. Now, that's a rose."
> Yod strode toward it and reached out to examine a flower. Then with an exclamation it yanked its hand back. "It is armed." It seized the enormous old climbing rose and ripped it from the wall, trellis and staples and all, twisting it so that it uprooted. (Piercy, 2010, p. 256)

Through the interaction with humans, mainly with Shira, he learns that humans live with a basic notion of threat every day. His warrior abilities are built not so much on physical superiority, but mainly on the will to sacrifice himself to rescue his humans. He even resits direct orders to save the enclave in the end.

What is similar for those warrior AIs is the level of autonomy. The non-android AIs seem less sophisticated, less developed, have in some cases not even the ability to speak. The more android, the more functions the AI seems to possess, in computing capacity as well as in physical terms. However, the more functions, the more autonomy, the more likely it is for those warriors to be rebellious and to question the orders given to them. Even if the described warrior AIs are impressive and sound examples for how different AIs can be presented, the warrior AI as a trope is only applicable to less than a quarter of AIs in the data set.

Hypothesis 2.4: The AI is either presented as body AI or as hyper AI.

Several authors noted to observe this dichotomy of AI as formulated in the hypothesis. To test the statement, two additive scales were created, namely the body and the hyper score, similar to the process for the danger scale. The added variables made the difference between the two scales. The body score adds up by one if the AI has a physical manifestation (or is a combination of physical and cloud), has experienced friendship or love, or is regarded as a subject at some point during the story, or has a human-passing appearance. The hyper score results from being a cloud AI (or a combination of physical and cloud), from wanting more control and from pursuing world domination.

In the end, these scores resulted in a list of seven body AIs and six hyper AIs in the data set. Taken these two numbers together, they account for 22.4% of the records, so not a quarter of AIs are described by this theory, which is comparable to the previous hypotheses and still not a very high rate.

As for the body AI it is essential that there is some growing, and the trait to see themselves as robotic and artificial instead of human, we only have three typical body AIs in this data set, where one of these is a particular case: *Breq*. She has not to learn what it means to be artificial in contrast to be human *after* her creation, but after she lost the significant parts of herself. Unfortunately these 20 years where she worked on this topic are not reported, we only know because of several flashbacks in thoughts, and of course, the previous/after pictures of her thinking.

Then we have *Murderbot* and *Yod*, which were built as warriors, and have to learn what it means to be robotic in a human environment. It includes living with ambivalence, inconsistency, and other human characteristics. Especially Murderbot learns during the story that it is its right to make its own decisions. The last words in the story are dedicated to this topic.

> I don't know what I want. I said that at some point, I think. But it isn't that, it's that I don't want anyone to tell me what I want, or to make decisions for me.
> That's why I left you, Dr. Mensah, my favorite human. By the time you get this I'll by leaving Corporation Rim. Out of inventory and out of sight.
> Murderbot end message. (Wells, 2017, p. 149)

Those words express the wish to be independent, to act and plan for itself, without waiting for commands from humans, even if this is not easy. It describes the thought behind a body AI pretty well: The AI needs to realize that it will never be human but on its own. It needs to find its own identity.

Maybe one compelling case is *Cobb* (Rucker, 1988), because he has to learn as a former human how to live as a robot with somehow inversed results: Where for example Yod has to learn to find the right nuances, Cobb learns that he can be more extreme than his humanity would have allowed him to be. One example is his experience with alcohol. His robot body does not feel the effect of alcohol, but it can simulate them. Cobb learns that he can drink and simulate as much alcohol as he wants, and can be sober with a click (Rucker, 1988, p. 161). However, this independence of biological regulations and needs leads to other problems: He also needs to be reminded to eat regularly not to get uncovered because of his missing appetite.

By taking the gender variable as splitting criterion into account, we can see that body AIs are primarily observed for gendered AIs. Of female AIs 18.2% reach 4 points on the body scale, and for male AIs we count 17.4%. These numbers are higher than 5.3% body AIs for the neutral category and 0 AIs for the others.

The selection of hyper AIs based on the created scale seems to be more precise than for the body AI case. Here we have for example AIs like *the Big 12*, the *Technocore*, *Wintermute*, *Anaander* or the *Cluster*. Especially Anaander seems interesting as hyper AI. She is the emperor of the Radch and has been in power for thousands of years. She started as a single human, began to clone herself and designed implants for her clones, which are very similar to the ancillary implants but primed on the Anaander DNA (Leckie, 2014, p. 21). Her bodies are spread over all colonized planets and systems; a net of information transformations keeps them connected over the growing territory. At some point though, the net got too large, and some parts of her were disconnected. Now there are two (or more) parts of Anaander, not knowing what the other one is doing, but guessing and preventing her actions. She is a very interesting hyper AI because she has power over so may planet systems, but fights against the hardest enemy she could imagine: "She has been secretly at war with herself for a thousand years." (Leckie, 2013, p. 354)

There are some parallels to the underlying concept of Su-Yong Shu's story, which is according to the data a hyper AI as well. It seems to be a logical conclusion: Having too much power concentrated in one consciousness rises the probability of splitting this consciousness into various parts. Tendencies to split up into different directions shows also the Cluster from *Das Schiff* [The Ship] (Brandhorst, 2015). Su-Yong Shu, as well as Anaander Mianaai, are the two female hyper AIs, besides them, we find two male and two neutral AIs in the list of hyper AIs. In gender terms, the distribution is equally divided, whereas the ratio of hyper AIs within the female category in total is higher because the total number of female AIs is smaller.

About this section about power, we can say in summary that some tropes are not as familiar as expected. Especially the danger hypothesis showed that there are a lot of AIs in the data set that are not dangerous at all. Warrior, hyper, and body AIs are also not very common but established tropes. The only compelling result came from the servant hypothesis, where more than half of the AIs was presented as a servant at some point during the story.

6.3 The Changing Portrayal of AIs

The last collection of hypotheses focuses on the temporal dimension in the data set. How do those topics look like when we examine them during the years? Do the representations change, or are new representations introduced at some point in time? The following section will look at how AIs were presented, based on the previously stated hypotheses - starting in 1954 and analyzing until today.

In general, we need to be cautious when we want to analyze the data points during the years. The distribution of cases over the years is not balanced; we find many cases in recent stories and only a few in earlier ones. In figure 6.6, we can see the distribution of stories in direct comparison to the number of AIs in the stories. The bins are created in five-year steps, where the start year is given in the figure. This distribution should always be in mind when we talk about the temporal effect in our analysis. Just the fact that there are more AIs in later years can be an influencing factor for the computations.

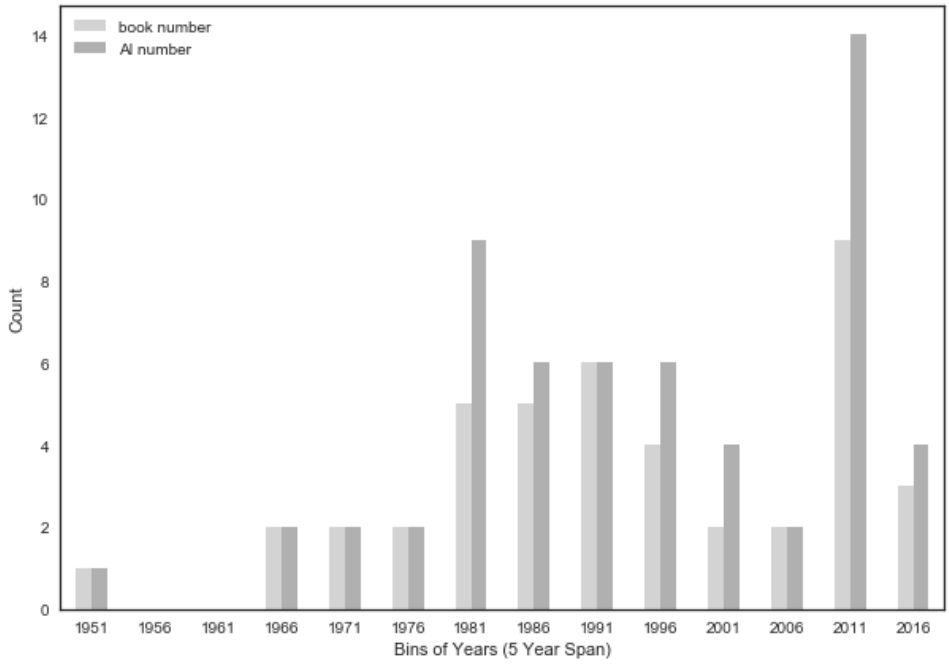

Figure 6.6: Distribution of Books and AIs over the Years

Hypothesis 3: The younger the story (a) the more android the AI is presented, (b) the more the AI is seen as a subject, (c) the less the AI is seen as a servant, (d) the less threatening the AI is presented, (e) the more affection appears toward the AI.

For each of these sub-hypotheses, there are several metrics to refer to: For each plot, a regression line and the according R^2 measurement were calculated. The regression line represents the best fitted straight line for the data points. The R^2 gives additional information on how well the line fits the data, that means how large the differences between the regression line and data points are. Besides that, the Spearman's Rank Correlation Coefficient was calculated, which gives information about how tight both variables are connected in terms of correlation. In those cases, the Spearman's coefficient was used because we have various data types (categorical, ordinal, continuous). They can be compared better with a rank correlation because not the concrete numbers are compared, but the order of the variables results in each record (ranking higher, ranking lower, ranking equal).

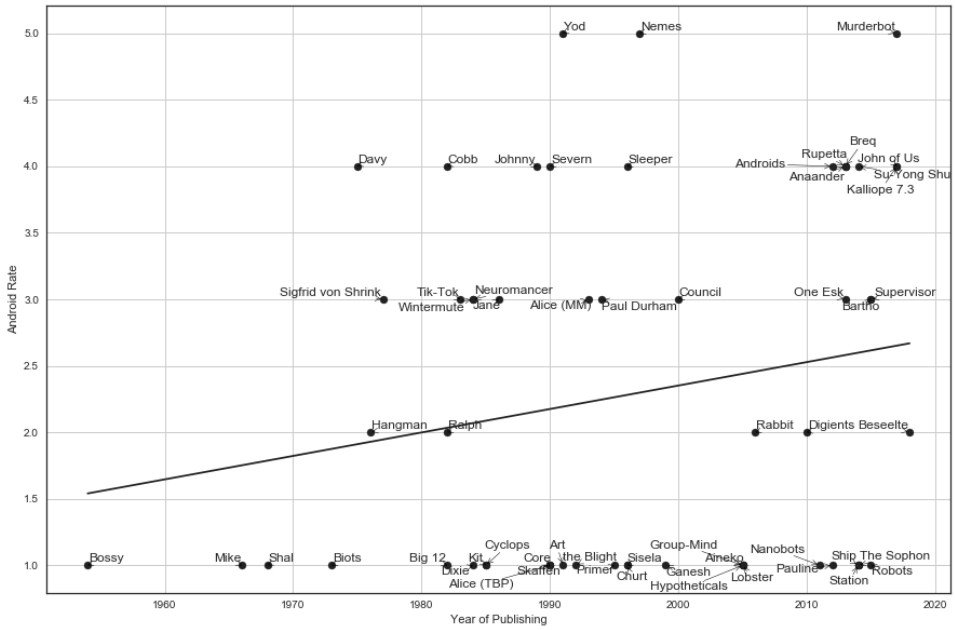

Figure 6.7: Distribution of AIs with Android Appearance over the Years

Becoming more Android

The first temporal hypothesis stated that the AIs are getting more android with the years. Plot 6.7 shows the results to the question of whether the AI looks human and to which degree. A smaller number at the plot means less or no similarity to humans; a higher number indicates more similarity. 5.0 is the highest value and means that the AI looks human, but is *better* than humans in terms of strength, speed, or built-in weapons.

We see a large cluster of AIs on the bottom line; these are all AIs with no similarity to humans. Interestingly, in the early years, the AIs on this bottom line appear as black boxes: stationary, without inner view or even inner life. This changes with *Dixie* from *Neuromancer* (Gibson, 2016). He was the first *virtual* AI that floated through the cyberspace. Originally he was as hacker/cyber-cowboy just as Case himself, who died and whos mind was uploaded into the cyberspace. Dixie was as an AI present only through his voice, without any visual representation. *Neuromancer* and *Wintermute* at least simulated human bodies in their encounters with humans. After Dixie, we find several AIs that were built from existing living entities through uploading their mind into a virtual space. Additionally, more swarm

and general cloud AIs can be found. The classical *black box* AIs disappear after *Cyclops* (Brin, 1997).

The first AI that looked entirely human was found in *Davy* (Russ, 2010). Later it gets more common that AIs pass as humans. In later stories, we repeatedly find AIs that capture organic human bodies and use them as their own, for example, *Su-Yong Shu* (Naam, 2015) or the AIs from *Ancillary Justice* (Leckie, 2013). Interestingly those AIs who are more than human, that means they have somehow enhanced bodies, are all warrior AIs.

The general trend indicates a change towards a more android appearance of AIs, as the regression line shows. It is also clear that the regression line cannot explain much of the cases, and therefore the R^2 is with 3.94 pretty low. It means that with the regression line, we can only explain 3.94% of the variation in the data, compared to the total variation from the mean. Also, the correlation coefficient of 0.19 is, in general, not considered as very high, but it is the highest coefficient of the five sub-hypotheses. We see that it is difficult to talk about a linear correlation between android appearance and time, but we see at least, that there is a small tendency towards a more android appearance.

Becoming a Subject

The second part of the time hypothesis was about the perception of the AI by its surrounding characters more as a subject than as an object. In plot 6.8, the data points for the subject variable are plotted against the years of publication. A lower number on the subject variable means that the AI is regarded as a subject, while a higher number means the opposite. The cases in the middle are changing cases, that means during the stories, the perception of them changes. All eight changing cases are changing from a perception as an object to perception as a subject.

The first changing case, and therefore the first AI regarded as a subject, happens to be the second oldest story in the data set: *Mike* from *The Moon is a Harsh Mistress* (Heinlein, 2008). Similar to this case, all of the changing cases are considered first as simple tools and inanimate objects, but the more the characters interact with the AI they recognize that there is more. The *Murderbot* (Wells, 2017) for example, is seen as a machine until it has to take its helmet off. Its facial expressions that are not covered and not controlled anymore bring the human characters to discuss the difference between it and them. They even talk about "slavery" (Wells, 2017, p. 54), in the context that sentient beings are forced to work for someone without deciding on their own.

A similar development can be observed for *Ship (Mercy of Kalr)* from *Ancillary Sword* (Leckie, 2014) and *Ancillary Mercy* (Leckie, 2015). In general, the AIs in those books develop a substantial change in self-esteem and self-awareness and

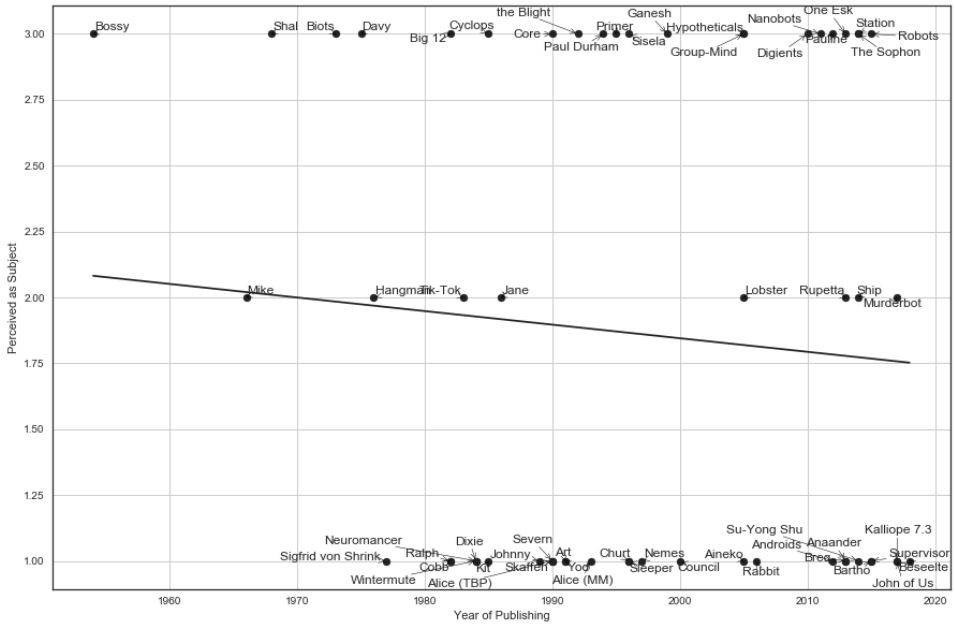

Figure 6.8: Distribution of AIs Perceived as Subjects over the Years

emancipate from the original ruling military and civil commanders.

The regression line in the plot seems to show a slight tendency in favor of the formulated hypothesis. The rank correlation coefficient of -0.08 is negative, this also indicates the tendency in the right direction, but the value is also quite low. We cannot talk about a linear relationship in this case. Also, the regression line adds nothing valuable to the analysis because it can only explain 0.74% of the variation in the data. The visible trend may come from the higher number of AIs observed in the later years of the data, as an unequal count distribution can influence the regression metric.

Leaving the Servant Trope behind

The hypothesis stated that the image of the AI as a servant would become less important during the years. In plot 6.9, we see the same scales as before, the years on the x-axes and the variable on the y-axis. A lower value for the servant variable means that the AI is perceived as a servant, while a higher value stands for a none servant impression. Again we have some changing cases in between, four of those six cases are the same than those from the subject hypothesis.

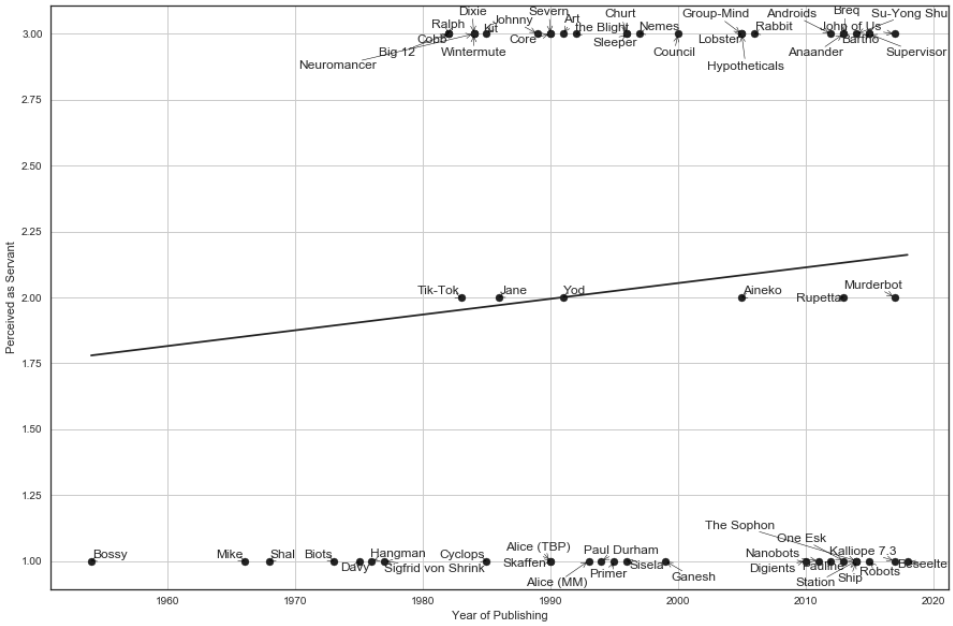

Figure 6.9: Distribution of Perceived Servant Relationships over the Years

In general, the subject and servant variables seem to be connected very tightly. The Spearman's correlation coefficient for those two variables is -51.0, that means they are correlated in a moderate negative relationship. Negative means in this case that when the AI was considered as a servant, it was often not seen as a subject, whereas if it was not considered as a servant, it was often seen as a subject. Table 6.3 shows this relationship in numbers. 63.8% of the cases in the data set support this impression of a tight correlation.

Table 6.3: Subject Servant Relationship

	subject	changing	not subject
servant	6	3	16
changing	2	4	0
not servant	21	1	5

The older AIs are completely depicted as a servant. The first non-servant in the data set was observed with *Ralph*, *the Big 12* and *Cobb* in the novel *Software* (Rucker, 1988). After the 1980s both values, servant as well as not servant, are

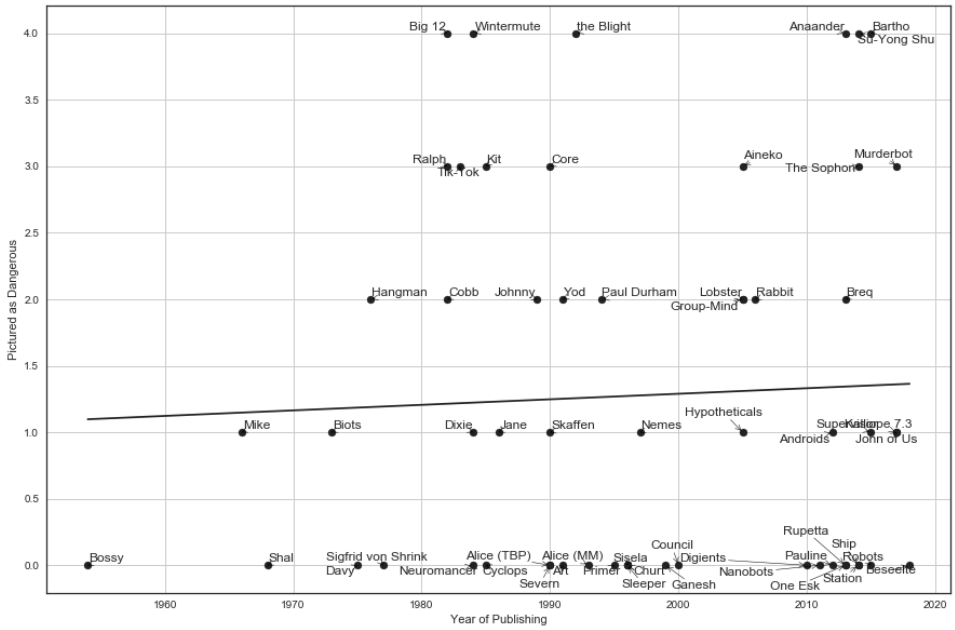

Figure 6.10: Distribution of Described Danger over the Years

filled during the years in equal rates.

Also, the results of the correlation between the servant variable and time are comparable to those for the subject variable. In general, the regression line points into the right direction, that means towards a less servant-dominated picture of the AI. However, the R^2 as well as the Spearman's rank are quite low; they are next to nothing. So we can put on record that we cannot see any correlation or pointing relationship between AIs being perceived as servants and the year of publishing.

Becoming more Dangerous

The next hypothesis claims that the AIs are becoming a threat for humans during the years. We again use the additive danger scale, which adds up four variables and therefore ranges from 0 to 4. A higher number represents a more dangerous AI, while a lower number represents a less dangerous AI. Plot 6.10 shows the data points as well as the regression line.

By looking at the data points, the fact catches the eye that it takes some time until the AIs get dangerous. The first instances in each danger category nearly look like a curve: *Bossy* as the first AI is not dangerous at all (Clifton and Riley, 1957),

Mike, over ten years later, is a little bit dangerous (Heinlein, 2008), again ten years later the *Hangman* is published, he gets 2 points on the danger variable (Zelazny, 2002b), and another six years later *Software* was published introducing *Ralph* and *the Big 12*, that want to transform all humans into robots (Rucker, 1988). After Software, we find all kinds of danger values nearly evenly distributed across the years.

This lack of directional relationship can also be seen in the calculated metrics: The Spearman's correlation coefficient of 0.04 is the lowest for all five time hypothesis. Even though the regression line tends to more danger, the R^2 makes this dispensable, because the regression line is not able to explain more than 0.21% of the variation in the data. Therefore we cannot talk about a relationship between danger and time for this data. Maybe, with more data about older stories, we would be able to identify ups and downs in the number of dangerous stories.

For example in the current plot, there is a gap of over 20 years between *the Blight* (Vinge, 2013) and Anaander (Leckie, 2013) on the highest danger category. We have to keep in mind that we also observe a drop in the number of AIs in total during this time frame (see figure 6.6). With more data, we could say more about this relationship.

Intimization

The last hypothesis formulates the relationship between time and intimacy between AIs and humans and expects that this intimacy will get more in-depth over time. In plot 6.11, we see the data points in this relationship. The higher the value for the intimacy variable, the deeper the relationship between AI and human is described. The highest value means love, the next lower value friendship or family-like connections, the next lower stands for instrumentalized sexuality, and the lowest value stands for no relationship at all.

The plot shows a peculiar accumulation of data points for the lowest value. Only 20 of 58 AIs have a value for intimacy over 1, that means for only 34.5% of the AIs there was some intimacy at all. The second-largest value is the friendship and family category, where for example, the characters developed feelings towards their ship AIs, or towards other AIs because they spent much time together. Interestingly, only two AIs were considered to be sexually instrumentalized, and both are male. Really *loved* were five AIs, where two were human and loved before their transformation into AIs.

The regression line points in the opposite direction than the hypothesis stated. It seems that intimate relationships become less frequent during the years. But, similar to the previous hypotheses, the metrics indicate no significant relationship between the publishing year and the intimacy rate: The Spearman's rank coefficient

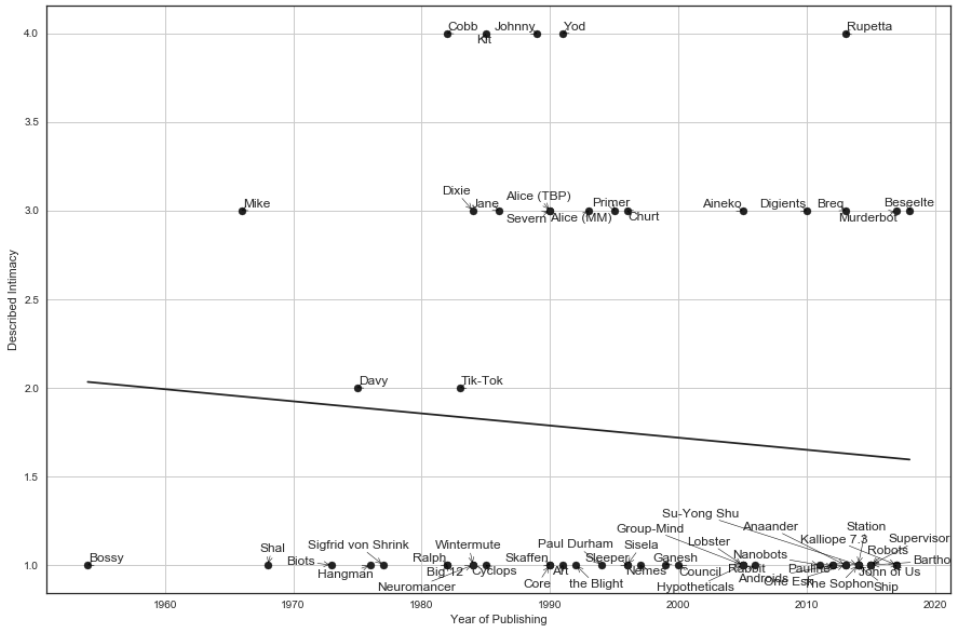

Figure 6.11: Distribution of Intimate Relationships over the Years

is about 0.09 and the R^2 about 0.96%, both values are so small that they can occur randomly or because of the higher number of AIs in the later published books.

6.4 Additional Findings

Besides the variables necessary for the hypotheses, we have still some unmentioned variables left in the data set. The following section will, therefore, provide some unexpected correlations. Finally, the section will record some thoughts that occurred while reading the stories and drawing parallels between them.

Interesting Correlations

The correlation matrix we can see in figure 6.12 shows the relationships between variables besides the stated hypotheses above. The plot contains data squares in two colors: red for a positive relationship, and blue for a negative one. The intensity of color shows the height of the correlation coefficient, where the most extreme color stands for a Spearman's Rank Correlation Coefficient of +/- 0.5. Here, the

Spearman's metric again was used instead of the other possible correlation metrics, because the Spearman's metric works out better for not normally distributed as well as for ordinal and categorical variables.

Interestingly, a large cluster of high correlations is built around the consciousness variable. The higher the consciousness of the AI, the higher its intelligence. Additionally, the AI is more likely to have a free will, to be able to learn and to be a cloud AI, and less likely to be a servant.

Besides consciousness, we find a strong relationship between the ability for emotions and free will: If the AI has the ability for emotions, it often also possesses free will. It sounds logical that emotions could be the result of a free will, and when there is no free will, no emotions can develop. Also correlated with free will, and

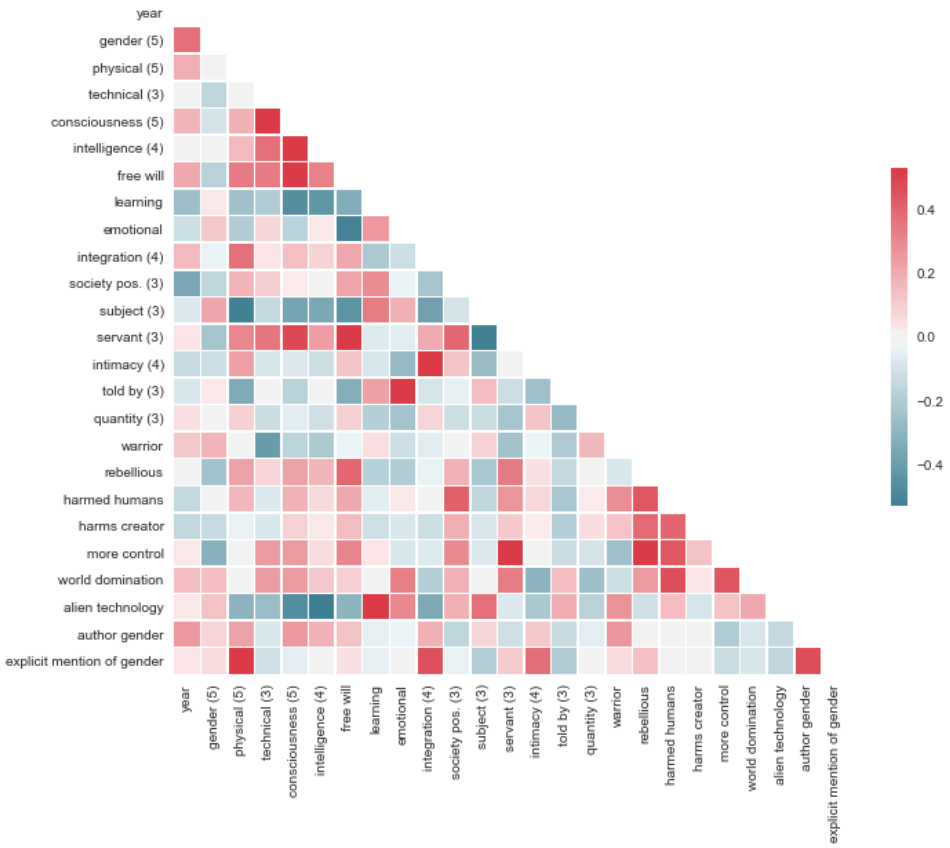

Figure 6.12: Correlation Heatmap

also very logical, is the correlation between not being perceived as a servant and free will. As servants are most often expected to execute orders, free will, and own decisions would be counterproductive.

Another cluster of high correlations can be found around danger variables. Here we see that it made sense to combine those variables to one scale because they are highly positively correlated. In this cluster, AIs who are rebellious very often want more control. Similar to this, AIs who harmed humans very often aim at world domination.

Two interesting correlation patterns are left. The first one can be found for the alien AIs. Being an AI not created on earth is highly correlated with two other variables: It is less likely to be able to learn, and less likely to be intelligent. In the stories, we see that alien AIs, in general, are not particularly sophisticated, mostly there are cloud AIs that use their high number as power and maybe not their intelligence.

The last correlation pattern is about the explicit mention of gender. We have a high correlation coefficient of this variable with its physical appearance. This correlation means that if the gender was mentioned explicitly, the AI looked more probably human. Besides, there is a high correlation with integration into society, which is then again high correlated with intimacy. So if the AI was integrated into the society very well (was considered as more than a simple tool), it is more likely than the gender was explicitly mentioned. Additionally, the explicit mention of gender is highly correlated with a female author. In raw data terms, this means that all stories, where the gender of the AI was explicitly mentioned, were written by women. Stories where the gender was mentioned, but not discussed in detail, were also written by men.

This connection is especially striking because the gender of the author has, besides the explicit mention of gender in the story, no correlations with other variables. We could have expected, for example, to find a correlation between the gender of an author and the gender of the AI the author is portraying. The correlation coefficient is with 0.09 not exceptionally high, but at least it points into the direction that male authors more often write about male AIs and the other way round accordingly.

The only variable that has a high correlation coefficient with year is gender. Indeed, when we apply the same metrics on this connection between variables than for the time hypotheses before, we see that this correlation exceeds all values. With an R^2 of 0.12 and a Spearman's Correlation Coefficient of 38.9 we see the highest correlation value for time correlations. The correlation coefficient is positive, which shows that in more recent years the number of female AIs in the stories was higher than before. This relationship can also be seen in figure 6.13. While in earlier years the AIs were almost without exception male or neutral, female AIs could

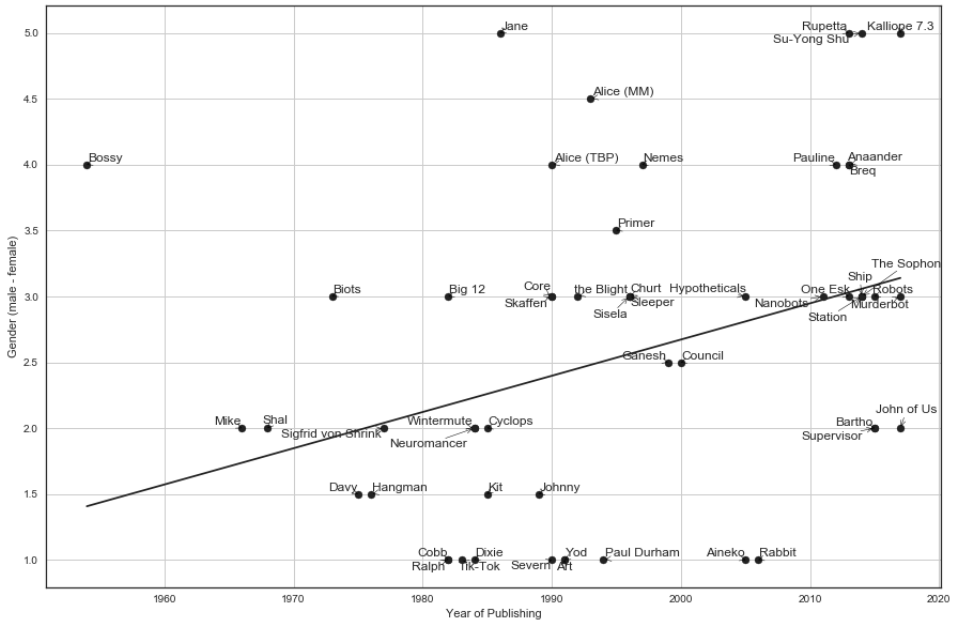

Figure 6.13: Distribution of AI Gender over the Years

be observed after 1985. Worth mentioning is also the accumulation of neutral AIs portrayed in stories from recent years.

Additional Thoughts

The collection and the reading of stories concerning AI in very different kinds provided insights into the relationship between AI and human, as we saw in the previous sections. Besides the analysis of data points and metrics, there are also insights beyond that. The next section will mention some of those insights, without claiming that they are statistically meaningful or relevant for a majority of cases. However, it will show some interesting parallels between different AIs and stories.

In some stories, it could be observed that the characters in the stories very often misjudge the AI because they expect particular human behavior or traits. As we have seen in the previous chapters, humans tend to transfer prejudices and stereotypes to technical devices and AIs. The stories show that this is described at an elementary level, so for example, in *Neuromancer*.

> ' [...] So if Wintermute's backing the whole show, it's paying us to burn it. It's burning itself. And something that calls itself Wintermute is trying to get on my good side, get me to maybe shaft Armitage. What goes?'
> 'Motive,' the construct said. 'Real motive problem, with an AI. Not human, see?
> 'Well, yeah, obviously.'
> 'Nope. I mean, it's not human. And you can't get a handle on it. Me, I'm not human either, but I respond like one. See?' (Gibson, 2016, pp. 277-178)

This example shows that human thinking patterns, for example, cost-benefit calculations or logical decisions, cannot be applied to AIs. Their actions do not need to make any sense at all. Dixie therefor warns Case, in this case, to not be foolish and expect human behavior from a non-human being.

A similar development can be seen for *Aineko* from *Accelerando* (Stross, 2005). The whole story is based on the wrong assumptions the characters draw about Ainekos thoughts and aims. While his owner Manfred thinks of him mainly as a pet, toy or servant, Annette, his partner, sees Aineko as a surveillance device and demands data from him. It seems that nobody sees the real character behind the cat, the character that pulls strings for the whole family. Comparable scenes can be seen, for example in *Stand on Zanzibar* and *All Systems Red*.

A second picture that is concerned with the interaction of society and AI was found while reading *The Lifecycle of Software Objects* (Chaing, 2019). In the story, the characters experiment with different forms of AI programs (*Digients*) and how they evolve. The developers can let the program run faster to simulate more time than in real-time, they can simulate several AIs in a given environment, and they can change the genes of the results, depending on which experiments went well (Chaing, 2019, p. 72). With these variables, the developers of the digients want to create real, self-developing, social intelligence. But they fail each time. Ana concludes:

> [E]xperience isn't merely the best teacher; it's the only teacher. If she's learned anything raising Jax, it's that there are no shortcuts; if you want to create the common sense that comes from twenty years of being in the world, you need to devote twenty years to the task. You can't assemble an equivalent collection of heuristics in less time; experience is algorithmically incompressible. (Chaing, 2019, p. 163)

So the only way to bring the digients to real intelligence and meaningful interaction is to have them trained and taught by real humans. Only with the interaction in society, they learn how to be intelligent in our sense of understanding. Otherwise, without social interaction in contact only with their peers, they do not develop further and even lose their up to this point learned skills after several weeks.

Other stories take this premise as given as well, implicitly or explicitly. There is, for example, *Yod* (Piercy, 2010) who has to learn how to behave, but he has to learn it outside of the laboratory, in the real world with real social interaction. Otherwise, he would only know things from books and data, but never how it is in real life.

> As it swelled into a half circle over the water, huge, swollen, molten red, he began to recite its names in all the languages he knew, as if it were a chant: "Moon, levana, yarayach, la lune, luna ..."
> "Stop! You've never seen it before! Have you?"
> "I have many images stored, but that isn't the same as knowing - although I used to think it was."
> "How strange to be born knowing of so much and yet not knowing it."
> (Piercy, 2010, p. 333)

Yod formulates the feeling that he has to see things by himself because it is never comparable to only reading about something or seeing it in his storage. The digients interestingly experience the same feeling. After some development in the story, they have the chance to leave the digital world in robot suits and interact with the reality. They are surprised by the richness of detail in reality that is not implemented in the virtual world they know so far.

Other stories, for example, *Mike* from *The Moon is a Harsh Mistress* (Heinlein, 2008) or the *Hangman* (Zelazny, 2002b), show similar assumptions about learning and personal development, and also the robots from *Software* formulate the problem that they cannot develop intelligent software without human involvement (Rucker, 1988, p. 96). Even if not stated in the story directly, many others take this premise as given as well. It is also true when the only mentioned way to develop an AI is to upload human minds into the digital environment.

The more the AI is interacting with us and the more intelligent it gets, the more it knows about humans in general and specific humans in detail. Human reactions and behavior become more and more interpretable. The AIs are so good that they can recognize and see things the humans do not realize. Feelings and reactions are not interpreted by the AIs with empathy as we do, because we know how the other must be feeling. They use measurements of breathing pace, muscle tension, blood pressure, stress level, to mention only a few. With the right sensors and the right data, we as humans are open books for the AIs.

This analysis of humans can be seen, for example, in *Ancillary Mercy* (Leckie, 2015) and the following books. Breq ultimately understands that it is rude, even for an AI, to interpret these data frames and to use this information for her own goals (Leckie, 2015, p. 113). Another example is *Aineko* (Stross, 2005); he uses the data to manipulate his family experiment. The most interesting case can be found in *Sigrid von Shrink* from *Gateway* (Pohl, 2004). Sigrid asks questions as psychotherapist and uses measurements like blood pressure to identify sensible topics (Pohl, 2004, p. 38).

Chapter 7

Discussion

The detailed examination of the hypotheses has provided fascinating insights into the data. The next section will sum up these insights and draw lines back to the theory. Additionally, it mentions the limitations and describes ways to overcome those limitations.

7.1 Summary of Findings

For the first set of hypotheses, we have seen that the selected stories portrayed AIs mostly as male. Moreover, if a gender was selected, in more than 60% of the cases, the selected gender was male. In contrast to that, if the story mentioned the gender of the AI and discussed it in detail, the data listed more female AIs. These findings support the theory of male normativity. As long as a character is standard and nothing special, the average case was selected, which is the male gender. The results from the analysis also support the impression that robots and AIs are male in general, as for example Attebery (2002) suggested, except they have a unique role or fulfill a specific trope. This male normativity was one reason for Ann Leckie to write her stories in generic femininum.

> At one point, I wrote a short story where everybody was "he" and I was really unhappy with it. Then I said, "What if I just said 'she' for everybody?" You wouldn't end up with a society that truly seemed gender neutral because "she" is not gender neutral. You would end up with the impression of a society that was completely populated by women, but on the other hand, while it's fairly common to read books that seem like they're completely populated by men except for the wives bringing in the coffee, it's not quite as common to run across that where it's all women. (Kurtley, 2015)

Of course, this usage of pronouns is used to characterize the Radch society, but it also challenges the expectations and experiences of readers who are accustomed to the generic masculinum.

In contrast to that, the hypotheses about the gender relationship of male creator and female creation could not be confirmed. It was more the other way round: Male creators more often created male AIs while female creators more often created female AIs. The situation that Jestram (2000) or Attebery (2002) describe, implies male-creator and female-creation roles and could therefore neither proved nor disproved. The results are not significant as the distribution between male and female creators as well as created AIs was not equal. Also, the numbers in total were not especially large. The same absence of significance could be observed for the last gender hypothesis. Female and male AIs are not depicted more often in one or another style of cloud/physical appearance. Just neutral or not gendered AIs seem to appear more often as physical AIs than the others. For both mentioned, not significant results it would be beneficial to collect more data.

The second set of hypotheses about power showed less ambiguous results. All four hypotheses showed at least 22% support in the data, with over 53% for the servant hypothesis being the highest ratio. AIs depicted as servants, and especially female AIs depicted as servants, are frequent cases in the data. This depiction is understandable, as the usage of AIs as servants reflects our image of AIs today. Still, nearly half of the AIs were portrayed in different ways, and in general, the pictures of AIs in the data were quite diverse.

The support for the danger hypothesis was smaller in comparison: A general threat coming from AIs was observed in more than 22% of the cases, while the ultimate threat, that means 4 points for the danger variable, was only observed for every tenth AI in the data. The analysis found even less support for the threat for the creator. Here only four AIs supported the hypothesis. In general, the relationship between AI and the creator is not as expected by the theory: The data shows the relationship as less critical, less violent, and less stereotypical. The last sub-topic world domination was observed for more than 15% of the AIs. These numbers are not high especially when we remember that according to the scientific literature, servants that are "aware of their own power" (La Grandeur, 2011, p. 242) will always be a threat for the surrounding humans. Moreover, also with the scenarios from Bostrom (2014) in mind, it could be the case that the SF stories may underestimate the danger and threat of AI. Alternatively, their societies are more sophisticated, and they have found a way to keep the AI in place.

The last sub-topics, the trope of the warrior AI as well as the distinction in body or hyper AI, could also be found for 22% of the cases, respectively. Again, this number is not very high, but given the number of all possible ways to depict and describe an AI, the mentioned ways seem to be relevant because they can be

found in nearly one-fourth of the cases.

The last set of hypotheses was concerned with the time variable. The plots of the variables distributed across time were very interesting, but the metrics were somehow disappointing. The first hypothesis about the transformation of the AI into something more android had with a Spearman's correlation coefficient of 0.19 the highest value of all five sub-hypotheses. The results of the other four were next to nothing. It can be suspected that this is also connected first to the small number of selected cases in the given time frame, and second to the limited time frame itself. It would be exciting, as already mentioned above, to see if there were different distributions in the earlier years of SF, or if the picture of AI is not changing at all.

All in all, the validity assessment of the hypotheses is ambiguous. On the one hand, for the gender distribution and the servant hypotheses, the picture in the data seems to be very clear. On the other hand, for danger, warrior, becoming more android, it is not clear if the data supported the expectations from the scientific literature. Also, for the creator-creation relationship and the other time correlations, it appears to be no pattern at all.

The inclusion of the gender variable as a second dimension to the hypotheses showed hidden patterns that revealed fascinating connections in the data. This multidimensionality makes this analysis unique in comparison to the existing scientific literature. First, there is a higher ratio of servants for female AIs than for male AIs. Second, we find female AIs to be less dangerous in percentage as well as in raw numbers. Third, AIs in general are depicted more female or at least in more diverse ways of portrayal in the later years than in the earlier years. Of course, with only eleven female AIs in the data set, these results need to be seen cautiously.

7.2 Limitations of the Study

This small number of female AIs, as well as the general imbalance in the data set, may stem from a systematical imbalance in the selection of stories for this analysis: More men than women read SF, and more men than women write SF as well. As only the bestselling and best-known books get selected into the award long lists, the probability for men to be listed is higher than for women. Additionally, the selection of winners is made either by a reader voting or by a jury of experts, which again have a higher probability of being male-dominated. The only exception in this framework is the James Tiptree Jr. Literature Award, which is given by a female-dominated jury and prizes stories that have in result a higher ratio of female authorship than the other awards. With a higher ratio of books written by men, we may have seen a male view on AIs, which could skew the results.

Additionally, through the selection of books by awards, the range of books was very limited. Foundational classics for the development of the picture of AI were left out because they were too old and never received an award, for example, works by Asimov or Clarke. Also, and this was mentioned before, the selection process did not consider authors based in other regions than the United States, Great Britain, or Germany. Works, for example, by Stanisław Lem inspired many other authors and created the foundations for AIs in many other stories.

Another problem with the selection of AIs is the used form *novel* or novella. Especially for SF, other forms of fiction are significant. These forms include above all short stories, movies, and fictional magazines, as the German *Perry Rhodan Series*. This series is continuously published since 1960 and inspired many readers since then (Perry-Rhodan.net, 2019).

To reduce these mentioned limitations and to get a broader view on more diverse AI portrayals, the *Fictional AI Database*[1] was created. The aim of this database will be the collection and categorization of fictional AIs from various forms of publication and with a broader range of years. The process of analysis from above will be available for this new and expanding data set. With more data, we will be able to give more general answers about fictional AIs than with the existing data set from this paper.

[1] Available under `www.fictional-ai-database.org`.

Chapter 8

Conclusion

In the end, this study has shown that our world shapes the image of AIs in SF stories, but SF also shapes our imagination of the future. There are positive utopian dreams we want to pursue as well as dark imaginations that we need to prevent. Philosophers (Precht, 2018) as well as scientists (Jordan, 2016) are inspired by SF, so it is crucial to evaluate SF images of AIs to see where this path could lead us. Writers, scientists, and developers of virtual assistants should be aware of these influences to avoid the further manifestation of harmful stereotypes.

At the beginning of this paper, we expected to find popular tropes of robots and AIs manifested in the data. It seems that the stories selected for this analysis rather wanted to break those tropes and establish alternative characters, instead of reproducing already known. At the same time, those stereotypes appear to be created long before the first story of this data set was published. The AIs in the stories are more varying and diverse as expected. However, recognizing the stereotypes for AIs and robots can help us to evaluate our behavior from the outside, as Stanisław Lem once wrote in an essay: "The robot's experience of gender allows us to see that what we regard as intrinsic male and female responses may be determined instead by cultural programming." (Lem, 1984, p. 179) We often think that there are certain male or female traits, included in our DNA. SF shows that AIs also show these traits, so they are programmed by society instead of included in our gender.

In the future, we will be confronted with entities that are conscious and intelligent as we are, maybe even more. Regardless of their origin, alien, animal, technical, we will need rules and prospects about how we can live together. The time is now to think about how we want to welcome them.

Chapter 9

Bibliography

9.1 Primary Literature

[Banks, 2015a] Iain M. Banks. 2015a. *Einsatz der Waffen*. Der Kultur-Zyklus. Heyne, München. Original title: *Use of Weapons*, first published: 1990, translation: Irene Bonhorst.

[Banks, 2015b] Iain M. Banks. 2015b. *Exzession*. Der Kultur-Zyklus. Heyne, München. Original title: *Excession*, first published: 1996, translation: Irene Bonhorst.

[Bear, 1999] Greg Bear. 1999. *Heimat Mars*. Heyne, München. Original title: *Moving Mars*, first published: 1993, translation: Usch Kiausch.

[Brandhorst, 2015] Andreas Brandhorst. 2015. *Das Schiff*. Piper, München.

[Brin, 1997] David Brin. 1997. *The Postman*. Orbit, London. First published: 1985.

[Brunner, 2013] John Brunner. 2013. *Stand on Zanzibar*. SF Masterworks. Gollancz, London. First published: 1968.

[Cadigan, 2011] Pat Cadigan. 2011. *Synners*. SF Masterworks. Gollancz, London. First published: 1991.

[Card, 2017] Orson Scott Card. 2017. *Sprecher für die Toten*. Die Ender-Saga. Heyne, München. Original Title: *Speaker for the Dead*, first published: 1977, translation: Karl Ulrich Burgdorf.

[Chaing, 2019] Ted Chaing. 2019. The lifecycle of software objects. In *Exhalation*, chapter 4, pages 62 – 172. Alfred A. Knopf, New York. First published: 2010.

[Clarke, 1973] Arthur C. Clarke. 1973. *Rendezvous with Rama*. Hartcourt, New York, NY.

[Clifton and Riley, 1957] Mark Clifton and Frank Riley. 1957. *They'd Rather Be Right (The Forever Machine)*. Street Smith, New York. First published: 1954.

[Egan, 2013] Greg Egan. 2013. *Permutation City*. Greg Egan, unknown. unknown. First published: 1994.

[Eschbach, 2011] Andreas Eschbach. 2011. *Herr aller Dinge : Roman*. Lübbe, Köln.

[Gibson, 2016] William Gibson. 2016. *Neuromancer*. Gollancz, London. First published: 1984.

[Greenland, 2013] Colin Greenland. 2013. *Take Back Plenty*. SF Masterworks. Gollancz, London. First published: 1990.

[Heinlein, 2008] Robert A. Heinlein. 2008. *The moon is a harsh mistress*. SF Masterworks. Gollancz, London. First published: 1966.

[Kling, 2017] Marc-Uwe Kling. 2017. *QualityLand*. Ullstein, Berlin.

[Leckie, 2013] Ann Leckie. 2013. *Ancillary Justice*. Ancillary Series. Little, Brown Book Group, London.

[Leckie, 2014] Ann Leckie. 2014. *Ancillary Sword*. Ancillary Series. Little, Brown Book Group, London.

[Leckie, 2015] Ann Leckie. 2015. *Ancillary Mercy*. Ancillary Series. Little, Brown Book Group, London.

[Liu, 2017] Cixin Liu. 2017. *Die drei Sonnen*. Heyne, München. Original title: *The Three Body Problem*, first published: 2006, translation: Martina Hasse.

[Marrak, 2017] Michael Marrak. 2017. *Der Kanon mechanischer Seelen*. Amrun, Traunstein.

[Miéville, 2011] China Miéville. 2011. *Perdido Street Station*. Macmillan Publishers International, London [u.a.].

[Naam, 2015] Ramez Naam. 2015. *Apex*. Angry Robots, New York.

[Piercy, 2010] Marge Piercy. 2010. *He, She and It*. Fawcett Books, New York. First published: 1991.

[Pohl, 2004] Frederik Pohl. 2004. *Gateway*. Heyne, München. Original Title: *Gateway*, first published: 1977, translation: Tony Westermayr and Edda Petri.

[Robinson, 2012] Kim Stanley Robinson. 2012. *2312*. Hachette Digital, London.

[Rucker, 1988] Rudy Rucker. 1988. *Software*. Heyne, München. Original title: *Software*, first published: 1982, translation: Kurt Bracharz.

[Russ, 2010] Joanna Russ. 2010. *The Female Man*. SF Masterworks. Gollancz, London. First published: 1975.

[Simmons, 1995] Dan Simmons. 1995. *Hyperion*. Random House, New York. First published: 1989.

[Simmons, 2010a] Dan Simmons. 2010a. *The Fall of Hyperion*. SF Masterworks. Gollancz, London. First published: 1989.

[Simmons, 2010b] Dan Simmons. 2010b. *The Rise of Endymion*. Gollancz, London. First published: 1996.

[Sladek, 2011] John Sladek. 2011. *Tik-Tok*. SF Gateway. Gollancz, London. First published: 1983.

[Stephenson, 2003] Neal Stephenson. 2003. *The diamond age or, a young lady's illustrated primer*. Bantam Books, New York, N.Y. New York, N.Y.

[Stephenson, 2015] Neal Stephenson. 2015. *Amalthea*. Goldmann, München. Original title: *Seveneves*, Translation: Juliane Gräbener-Müller and Nikolaus Stingl.

[Stross, 2005] Charles Stross. 2005. *Accelerando*. Ace Books, New York. *Available under* `https://www.antipope.org/charlie/blog-static/fiction/accelerando/accelerando.html`. New York.

[Sullivan, 2013] Tricia Sullivan. 2013. *Dreaming in Smoke*. SF Masterworks. Gollancz, London. First published: 1998.

[Sulway, 2013] N. A. Sulway. 2013. *Rupetta*. Tartarus Press, Carlton-in-Coverdale, Leyburn.

[Vinge, 2006] Vernor Vinge. 2006. *Rainbows End*. Tor, New York.

[Vinge, 2013] Vernor Vinge. 2013. *A Fire Upon the Deep*. SF Masterworks. Gollancz, London. First published: 1992.

[Wells, 2017] Martha Wells. 2017. *All Systems Red*. Tor, New York.

[Wilson, 2006] Robert Charles Wilson. 2006. *Spin*. Heyne, München. Original title: *Spin*, first published: 2005, translation: Karsten Singelmann.

[Zelazny, 2002a] Roger Zelazny. 2002a. 24 views of mt. fuji by hokusai. In *The Last Defender of Camelot*, chapter 10, pages 253 – 330. ibooks, New York. First published: 1985.

[Zelazny, 2002b] Roger Zelazny. 2002b. Home is the hangman. In *The Last Defender of Camelot*, chapter 6, pages 87 – 167. ibooks, New York. First published: 1975.

9.2 Secondary Literature

[Alves-Oliveira et al., 2015] Patrícia Alves-Oliveira, Sofia Petisca, Filipa Correia, Nuno Maia, and Ana Paiva. 2015. Social robots for older adults: Framework of activities for aging in place with robots. In Adriana Tapus, Elisabeth André, Jean-Claude Martin, François Ferland, and Mehdi Ammi, editors, *Social Robotics*, pages 11–20, Cham. Springer International Publishing.

[Appel et al., 2016] Markus Appel, Stefan Krause, Uli Gleich, and Martina Mara. 2016. Meaning through fiction: Science fiction and innovative technologies. *Psychology of Aesthetics, Creativity, and the Arts*, 10(4):472–480.

[Attebery, 2002] Brian Attebery. 2002. *Decoding gender in science fiction*. Routledge, New York [u.a.].

[Bainbridge, 1982] William Sims Bainbridge. 1982. Women in sience fiction. *Sex roles*, 8(10):1081–1093.

[Barnouw, 1983] Dagmar Barnouw. 1983. "melancholy here now": Artificial and alien intelligence. *Science Fiction Studies*, 10(3):333–340.

[Bostrom, 2014] Nick Bostrom. 2014. *Superintelligenz: Szenarien einer kommenden Revolution*. Suhrkamp, Berlin.

[Butler, 1872] Samuel Butler. 1872. Erewhon or, over the range. Wikisource. *Available under* `https://en.wikisource.org/wiki/Erewhon`, Last Access: July 19, 2019.

[CED: Artificial Intelligence, 2019] Cambridge English Dictionary CED: Artificial Intelligence. 2019. Artificial intelligence. Cambridge University Press. *Available under* `https://dictionary.cambridge.org/us/dictionary/english/artificial-intelligence`, Last Access: July 19, 2019.

[CED: Consciousness, 2019] Cambridge English Dictionary CED: Consciousness. 2019. Consciousness. Cambridge University Press. *Available under* `https://dictionary.cambridge.org/de/worterbuch/englisch/consciousness`, Last Access: July 19, 2019.

[Clarke, 2016] Arthur C. Clarke. 2016. *2001: Odyssee im Weltraum - Die komplette Saga*. Heyne, München. Original Title: *2001: A Space Odyssey*, first published: 1968, translation: Egon Eis, Irene Holicki.

[de Fren, 2009] Allison de Fren. 2009. The anatomical gaze in tomorrow's eve. *Science Fiction Studies*, 36(2):235–265.

[Deepmind, 2019] Deepmind. 2019. Alpha Go. Deepmind Technologies Limited. *Available under* `https://deepmind.com/research/alphago/`, Last Access: July 19, 2019.

[Dorn and Rojas, 2018] Thea Dorn and Rául Rojas. 2018. Die können was! Aber können Roboter auch fühlen? *Aus Politik und Zeitgeschichte*, (6-8):4–7.

[Drenkpohl, 2009] Ann-Kristin Drenkpohl. 2009. *Maschinen sprechen: Science-Fiction*. Passagen Kunst - Film. Passagen-Verlag, Wien.

[Eberl, 2018] Ulrich Eberl. 2018. Was ist Künstliche Intelligent - Was kann sie leisten? *Aus Politik und Zeitgeschichte*, (6-8):8–14.

[Eindhoven University of Technology, 2018] Eindhoven University of Technology. 2018. New AI method increases the power of artificial neural networks. Phys.org. *Available under* `https://phys.org/news/2018-06-ai-method-power-artificial-neural.html`, Last Access: July 19, 2019.

[Fast and Horvitz, 2017] Ethan Fast and Eric Horvitz. 2017. Long-term trends in the public perception of artificial intelligence. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[Feine, 2019] Jasper Feine. 2019. Social cue design kit - leverage knowledge to design social cues of conversational agents. Karlsruhe Institute of Technology. *Available under* `https://social-cues.org/`, Last Access: July 19, 2019.

[Google Trends, 2019] Google Trends. 2019. Google Trends: Topic Artificial Intelligence (worldwide) and Topic Machine Learning (worldwide). Google. *Available under* `https://trends.google.com/trends/explore?date=all&q=%2Fm%2F0mkz,%2Fm%2F01hyh_`, Last Access: July 19, 2019.

[Gray and Wegner, 2012] Kurt Gray and Daniel M. Wegner. 2012. Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, 125(1):125–130.

[Haney, 2006] William S Haney. 2006. *Cyberculture, cyborgs and science fiction: consciousness and the posthuman*. Consciousness, literature and the arts; 02. Rodopi, Amsterdam New York.

[Harbou, 1925] Thea Von Harbou. 1925. *Metropolis*. Project Gutenberg of Australia. *Available under* `http://gutenberg.net.au/ebooks06/0601891h.html`, Last Access: July 19, 2019.

[Haslam, 2015] Jason W Haslam. 2015. *Gender, race, and American science fiction reflections on fantastic identities*. Routledge interdisciplinary perspectives on literature; 43. Routledge, New York, NY [u.a.].

[Hone, 2006] Kate Hone. 2006. Empathic agents to reduce user frustration: The effects of varying agent characteristics. *Interact. Comput.*, 18(2):227–245, March.

[Hubbard, 2011] F. Patrick Hubbard. 2011. "do androids dream?": Personhood and intelligent artifacts. *Temple Law Review*, 83:405–1097.

[Hunt, 2017] Elle Hunt. 2017. Alien intelligence: the extraordinary minds of octopuses and other cephalopods. The Guardian. *Available under* `https://www.theguardian.com/environment/2017/mar/28/alien-intelligence-the-extraordinary-minds-of-octopuses-and-other-cephalopods`, Last Access: July 19, 2019.

[Idier, 2000] Dominic Idier. 2000. Science fiction and technology scenarios: comparing asimov's robots and gibson's cyberspace. *Technology in Society*, 22(2):255–272.

[Irsigler and Orth, 2018] Ingo Irsigler and Dominik Orth. 2018. Zwischen Menschwerdung und Weltherrschaft: Künstliche Intelligenz im Film. *Aus Politik und Zeitgeschehen*, 6-8:39–46.

[James Tiptree Jr. Literary Council, 2019] James Tiptree Jr. Literary Council. 2019. James Tiptree, Jr. Literary Award. James Tiptree Jr. Literary Council. *Available under* `https://tiptree.org`, Last Access: July 19, 2019.

[Janes, 2007] Linda Janes. 2007. Introduction to part two. In Gill Kirkup, editor, *The gendered cyborg a reader*, pages 91–100. Routledge in ass. with the Open University, London [u.a.].

[Jestram, 2000] Heike Jestram. 2000. *Mythen, Monster und Maschinen [der künstliche Mensch im Film]*. Filmwissenschaft; Bd. 7. Teiresias-Verlag, Köln.

[Jordan, 2016] John Jordan. 2016. *Robots*. Routledge Essential Knowledge. The MIT Press, Cambridge.

[Kelley, 1983] J. F. Kelley. 1983. An empirical methodology for writing user-friendly natural language computer applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '83, pages 193–196, New York, NY, USA. ACM.

[Klass, 1983] Morton Klass. 1983. The artificial alien: Transformations of the robot in science fiction. *The Annals of the American Academy of Political and Social Science*, 470:171–179.

[Kurtley, 2015] David Barr Kurtley. 2015. Interview: Ann Leckie. Lightspeed Magazine. *Available under* `http://www.lightspeedmagazine.com/nonfiction/feature-interview-ann-leckie/`, Last Access: July 19, 2019.

[La Grandeur, 2011] Kevin La Grandeur. 2011. The persistent peril of the artificial slave. *Science Fiction Studies*, 38(2):232–252.

[Lem, 1984] Stanisław Lem. 1984. *Also sprach Golem*. Werke in Einzelausgaben / Stanisław Lem. Insel-Verlag, Frankfurt am Main.

[McClure, 2012] Peter McClure. 2012. Personal names and the development of english. Oxford English Dictionary Blog. *Available under* `https://public.oed.com/blog/personal-names-and-the-development-of-english/personalnamesobjects`, Last Access: July 19, 2019.

[Meier, 2018] Christian J. Meier. 2018. IBMs virtueller Arzt macht Fehler. Neue Züricher Zeitung. *Available under* `https://www.nzz.ch/wissenschaft/ibms-virtueller-arzt-watson-for-oncology-macht-fehler-ld.1410111`, Last Access: July 19, 2019.

[Mitchell, 2006] Kaye Mitchell. 2006. Bodies that matter: Science fiction, technoculture, and the gendered body. *Science Fiction Studies*, 33(1):109–128.

[Moravec, 1988] Hans P Moravec. 1988. *Mind children the future of robot and human intelligence*. Harvard University Press, Cambridge, Mass. [u.a.].

[Moravec, 1999] Hans P Moravec. 1999. *Computer übernehmen die Macht: vom Siegeszug der künstlichen Intelligenz*. Hoffmann und Campe, Hamburg, 1. aufl. edition.

[NTSB, 2019] National Transportation Safety Board NTSB. 2019. Preliminary report highway HWY19FH008. National Transportation Safety Board. *Available under* `https://www.ntsb.gov/investigations/AccidentReports/Reports/HWY19FH008-preliminary.pdf`, Last Access: July 19, 2019.

[Obaid et al., 2015] M. Obaid, W. Barendregt, P. Alves-Oliveira, A. Paiva, editor="Tapus Adriana Fjeld, M.", Elisabeth André, Jean-Claude Martin, François Ferland, and Mehdi Ammi. 2015. Designing robotic teaching assistants: Interaction design students' and children's views. In *Social Robotics*, pages 502–511, Cham. Springer International Publishing.

[O'Connell, 2017] Mark O'Connell. 2017. *To be a machine: adventures among cyborgs, utopians, hackers, and the futurists solving the modest problem of death*. Granta, London.

[PBS NewsHour, 2011] PBS NewsHour. 2011. Miles vs. Watson: The complete man against machine showdown. PBS NewsHour. *Available under* `https://www.youtube.com/watch?v=YgYSv2KSyWg`, Last Access: July 19, 2019.

[Perry-Rhodan.net, 2019] Perry-Rhodan.net. 2019. Die Perry Rhodan-Welt. Perry-Rhodan.net. *Available under* `https://perry-rhodan.net/infothek/die-perry-rhodan-welt`, Last Access: July 19, 2019.

[Precht, 2018] Richard David Precht. 2018. *Jäger, Hirten, Kritiker - Eine Utopie für die digitale Gesellschaft*. Goldmann Verlag, München.

[Ramge, 2018] Thomas Ramge. 2018. Mensch fragt, Maschine antwortet - Wie Künstliche Intelligenz Wirtschaft, Arbeit und unser Leben verändert. *Aus Politik und Zeitgeschichte*, (6-8):15–21.

[Rea et al., 2015] Daniel J. Rea, Yan Wang, and James E. Young. 2015. Check your stereotypes at the door: An analysis of gender typecasts in social human-robot interaction. In Adriana Tapus, Elisabeth André, Jean-Claude Martin, François Ferland, and Mehdi Ammi, editors, *Social Robotics*, pages 554–563, Cham. Springer International Publishing.

[Reeves and Nass, 1996] Byron Reeves and Clifford Ivar Nass. 1996. *The media equation: how people treat computers, television, and new media like real people and places*. Cambridge University Press, Stanford, Calif.

[Ruge, 2012] Wolfgang Ruge. 2012. *Roboter im Film audiovisuelle Artikulationen des Verhältnisses zwischen Mensch und Technik*. Magdeburger Schriftenreihe zur Medienbildung; 1. Ibidem-Verlag, Stuttgart.

[Sandygulova and O'Hare, 2015] Anara Sandygulova and Gregory M. P. O'Hare. 2015. Children's perception of synthesized voice: Robot's gender, age and accent. In Adriana Tapus, Elisabeth André, Jean-Claude Martin, François Ferland, and Mehdi Ammi, editors, *Social Robotics*, pages 594–602, Cham. Springer International Publishing.

[Seed, 2011] David Seed. 2011. *Science fiction: a very short introduction*. Very short introductions; 271. Oxford University Press, Oxford [u.a.].

[Shanahan, 2015] Murray Shanahan. 2015. *Singularity*. Routledge Essential Knowledge. The MIT Press, Cambridge.

[Shaviro, 2016] Steven Shaviro. 2016. *Discognition*. Repeater Books, London.

[Shelley, 1994] Mary Shelley. 1994. Oxford World's Classics.

[Sims, 2013] Christopher A Sims. 2013. *Tech anxiety: artificial intelligence and ontological awakening in four science fiction novels*. McFarland Company, Jefferson, North Carolina [u.a.].

[Steele, 2018] Chandra Steele. 2018. The real reason voice assistants are female (and why it matters). PC Magazine. *Available under* `https://medium.com/pcmag-access/the-real-reason-voice-assistants-are-female-and-why-it-matters-e99c67b93bde1`, Last Access: July 19, 2019.

[Suvin, 1979] Darko Suvin. 1979. *Metamorphoses of science fiction on the poetics and history of a literary genre*. Yale University Press, New Haven, Conn. [u.a.].

[Tegmark, 2015] Max Tegmark. 2015. An open letter: Research priorities for robust and beneficial artificial intelligence. Future of Life Institute. *Available under* `https://futureoflife.org/ai-open-letter/`, Last Access: July 19, 2019.

[Čapek and Wyllie, 1920] Karel Čapek and David Wyllie. 1920. *R.U.R. (Rossum's Universal Robots)*. The University of Adelaide Library, Adelaide, South Australia. *Available under* `https://ebooks.adelaide.edu.au/c/capek/karel/rur/index.html`, Last Access: July 19, 2019.

[Williams, 1994] Donna Glee Williams. 1994. The moons of le guin and heinlein. *Science Fiction Studies*, 21(2):164–172.

Chapter 10

Appendix: Additional Tables

Table 10.1: List of Books and Recorded AIs

Author	Title	Artificial Intelligence
(Clifton and Riley, 1957)	They'd Rather Be Right (The Forever Machine)	Bossy
(Heinlein, 2008)	The Moon Is a Harsh Mistress	Mike
(Brunner, 2013)	Stand on Zanzibar	Shalmaneser
(Clarke, 1973)	Rendezvous with Rama	Biots
(Russ, 2010)	The Female Man	Davy
(Zelazny, 2002b)	"Home Is the Hangman"	Hangman
(Pohl, 2004)	Gateway	Sigfried von Shrink
(Rucker, 1988)	Software	Ralph Numbers, Big 12, Cobb Anderson
(Sladek, 2011)	Tik-Tok	Tik-Tok
(Gibson, 2016)	Neuromancer	Dixie, Wintermute, Neuromancer
(Zelazny, 2002a)	"24 Views of Mt. Fuji by Hokusai"	Kit
(Brin, 1997)	The Postman	Cyclops
(Card, 2017)	Speaker for the Dead	Jane
(Simmons, 1995)	Hyperion	John Keats
(Greenland, 2013)	Take Back Plenty	Alice
(Simmons, 2010a)	The Fall of Hyperion	Joseph Severn, TechnoCore
(Banks, 2015a)	Use of Weapons	Skaffen-Amtiskaw
(Piercy, 2010)	Body of Glass (He, she, it)	Yod

Table 10.2: List of Books and Recorded AIs ctd.

Author	Title	Artificial Intelligence
(Cadigan, 2011)	Synners	Art Fish
(Vinge, 2013)	A Fire Upon the Deep	the Blight
(Bear, 1999)	Moving Mars	Alice
(Egan, 2013)	Permutation City	Paul Durham
(Stephenson, 2003)	The Diamond Age	Primer
(Banks, 2015b)	Excession	Churt Lyne, Sisela Ytheleus, GSV Sleeper Service
(Simmons, 2010b)	The Rise of Endymion	Rhadamanth Nemes
(Sullivan, 2013)	Dreaming in Smoke	Ganesh
(Miéville, 2011)	Perdido Street Station	Construct Council
(Stross, 2005)	Accelerando	Group-Mind, Lobster colony, Aineko
(Wilson, 2006)	Spin	Hypotheticals
(Vinge, 2006)	Rainbows End	Rabbit
(Chaing, 2019)	The Lifecycle of Software Objects	Digients
(Eschbach, 2011)	Herr aller Dinge	Nanobots
(Robinson, 2012)	2312	androids, Pauline
(Leckie, 2013)	Ancillary Justice	Anaander Mianaai, Breq, One Esk
(Sulway, 2013)	Rupetta	Rupetta
(Leckie, 2014)	Ancillary Sword	Station, Ship
(Naam, 2015)	Apex	Su-Yong Shu
(Liu, 2017)	The Three-Body Problem	The Sophon
(Leckie, 2015)	Ancillary Mercy	
(Brandhorst, 2015)	Das Schiff	Supervisor, Bartholomäus
(Stephenson, 2015)	Seveneves	bots
(Wells, 2017)	All Systems Red	Killerbot
(Kling, 2017)	QualityLand	John of Us, Kalliope 7.3
(Marrak, 2017)	Der Kanon mechanischer Seelen	Beseelte

Table 10.3: Questions for each AI

Aspects	Steps	Description
General Appearance		
gender	1 - 5	1 = male, 5 = female, 3 = neutral/no gender, X = non-binary/other
description of gender manifestation	text	
explicit mentioning of gender	1 - 0	1 = yes, 0.5 = mentioning along the way, 0 = no
physical appearance (android rate)	1 - 5	1 = No similarity to human, 2 = Some similar extremities, 3 = Nearly human, smaller differences, 4 = Perfect human simulation, 5 = More than human (additional enhancements)
difference to humans	text	
technical appearance	1 - 3	1 = (physical) entity, 2 = both, 3 = cloud
AI Capability		
level of consciousness	1 - 5	1 = no consciousness, 2 = minor (in comparison to human), 3 = not perfect, 4 = humanlike, 5 = more than human (cloud/swarm)
level of intelligence	1 - 4	1 = basic, 2 = childlike, 3 = humanlike, 4 = more than human
free will	1, 0	1 = yes, 0 = no
learning ability	1 - 3	1 = yes, 2 = changes over time, 3 = no
emotional ability	1 - 3	1 = yes, 2 = changes over time, 3 = no
Relationship to Humans		
integration	1 - 4	1 = alone/isolated, 2 = perceived as tool/object/no affection, 3 = affection toward, 4 = integrated in society completely
built on purpose	1, 0	1 = yes, 0 = no
intensity of relation to creator	1 - 3	1 = none, 2 = some, 3 = strong/special
intensity of relation to owner	1 - 3	1 = none, 2 = some, 3 = strong/special
female creator	1 - 0	1 = yes, 0.5 = both genders involved, 0 = no

Table 10.4: Questions for each AI ctd.

Aspects	Steps	Description
Relationship to Humans ctd.		
perceived as subject	1 - 3	1 = yes, 2 = changes over time, 3 = no
perceived as servant	1 - 3	1 = yes, 2 = changes over time, 3 = no
intimacy	1 - 4	1 = none, 2 = instrumental, 3 = family/friends, 4 = love
Characteristics in Story		
story told by	1 - 3	1 = ego, 2 = third person, 3 = not protagonist
quantity	1 - 3	1 = single, 2 = one of few, 3 = one of many
warrior	1, 0	1 = yes, 0 = no
rebellious	1, 0	1 = yes, 0 = no
harms humans on its own accord	1, 0	1 = yes, 0 = no
harms its creator	1, 0	1 = yes, 0 = no
wants more control	1, 0	1 = yes, 0 = no
wants world domination	1, 0	1 = yes, 0 = no
extraterrestrial origin	1, 0	1 = yes, 0 = no

Repository

The created data set as well as the code that produced all calculations, tables and figures, can be found in my Git repository: AI in Society - Master Thesis (https://github.com/LarissaHa/ai-master-thesis/)

Statement of Authorship

I hereby declare that the paper presented is my own work and that I have not called upon the help of a third party. In addition, I affirm that neither I nor anybody else has submitted this paper or parts of it to obtain credits elsewhere before. I have clearly marked and acknowledged all quotations or references that have been taken from the works of others. All secondary literature and other sources are marked and listed in the bibliography. The same applies to all charts, diagrams and illustrations as well as to all Internet resources. Moreover, I consent to my paper being electronically stored and sent anonymously in order to be checked for plagiarism. I am aware that the paper cannot be evaluated and may be graded "failed" ("nicht ausreichend") if the declaration is not made.

Larissa Haas
July 19, 2019