universität wien

DIPLOMARBEIT / DIPLOMA THESIS

Titel der Diplomarbeit / Title of the Diploma Thesis

„Artificial Intelligence in Science Fiction Literature - Could Humans be a Form of Artificial Intelligence?“

verfasst von / submitted by

Katharina Linemayr

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Magistra der Philosophie (Mag. Phil)

Wien, 2017/ Vienna, 2017

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 190 344 299
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Lehramtsstudium UF Englisch UF Psychologie und Philosophie
Betreut von / Supervisor:	Ao. Univ.-Prof. Mag. Dr. Eva Zettelmann

Contents

1 Introduction

Science fiction literature has been inspiring its readers to think in new dimensions since its very beginning. It is known to present visions of a future, which might yet dawn for mankind. However, science fiction does not claim to make accurate predictions, but should rather be seen as an "embodied thought experiment" (Seed 2). In these thought experiments the impossible often becomes possible. Technology is typically far more advanced than in the world we know today; space travel is taken as self-evident, people live on other planets, even in other galaxies. Such futuristic visions have also been adopted by two of the genre's most renowned authors: Arthur C. Clarke and Isaac Asimov. Both have shaped science fiction literature considerably. This thesis is concerned with *2001: A Space Odyssey* by Clarke and *I, Robot* by Asimov.

2001: A Space Odyssey's story spans from the very beginning of mankind's evolution from prehistoric times, when man-apes conquered the world, to an age of space travel. There it follows the journey of two astronauts and their highly intelligent on-board computer HAL. The novel ends with the birth of a star child, which represents humanity's next step in its evolution.

I, Robot comprises nine short stories about a future in which humans use robots as their servants. In each of the stories one particular robot is described by Susan Calvin, a robopsychologist, who has studied them throughout her life. The robots evolve from rather simplistic silent servants to creatures more intelligent than human beings, which take over the rule of the world.

Both of these novels, therefore, describe non-human intelligent creatures as protagonists in a futuristic world. HAL, the on-board computer, is able to speak and think and becomes a threat to the human crew as he tries to take over the spaceship and fulfil its mission

alone. In Asimov's work, the robots are also highly intelligent creatures, as they carry out important jobs on mining stations or in space. Such intelligent life forms are classified as artificial intelligence by modern science and are a much researched topic. Artificial intelligence in its perfection mimics human intelligence exactly (Friedenberg 5). If it is possible to create a machine, which comes close to human intelligence in the real world remains to be seen. However, in the novel's realities they are pictured as part of everyday life.

This thesis therefore is concerned with the intelligence of the non-human protagonists of the two novels in question. If one thinks of an intelligent computer or robot, one might imagine a deeply rational and analytic being. However, both HAL and the robots show signs of emotional reactions as well as rational thought throughout their stories. An emotional mind is normally associated with humans rather than machines, which raises the question whether there is something more to intelligence than one sees at the first glance. Artificial intelligent creatures might necessarily have to be emotional as well as rational beings. If this is the case, one might ask whether there really is a difference between human beings and creatures such as HAL and the robots. Clarke and Asimov could mean for them to be comparable to humans. If so, one can ask whether the term 'artificial intelligence' is a necessary label.

Hence, this thesis will analyse HAL and the robots in detail and will compare their intelligence to human intelligence to find out whether the two are similar or, in fact, the same.

2 Defining intelligences

2.1 Artificial intelligence

Regarding the two novels used in this paper, Arthur C. Clarke's *2001: A Space Odyssey*, further referred to as ASO, and Isaac Asimov's *I, Robot*, referred to as IR, one might think that the technological creatures portrayed within them have little in common between them. ASO depicts HAL as a spaceship's computer which runs amok, when threatened with a switch-off. IR features various connected short stories telling an evolutionary tale of robots, which cause a series of problems for human beings, until they finally reach a point of higher intelligence than them. However, as different as these creature's stories might seem, they can both be defined as beings of artificial intelligence, commonly abbreviated as AI. In order to establish this linking point, AI shall be defined, analysed and applied to the creatures in science fiction literature featured in this thesis.

Artificial intelligence is defined by the *Oxford Dictionary of Science Fiction* as "the name for the science of creating 'intelligent' computer programs, a sentient, self-aware computer or computer program". To go into more detail, two versions of AI can be defined, namely "classical AI" and "behavioural or interactionist AI" (Mateas 9). Mateas describes classical AI focusing on mental factors, while interactionist AI concentrates on the interaction of "embodied agents" (9). This means that classical AI describes the importance of internal processes in order for a machine to be intelligent but interactionist AI depends on a body to perform the intelligent action (Mateas 10). The core idea behind classical AI is said "to view mind as an abstract process, something that is not necessarily tied to the contingencies of human brains and bodies, but can rather be abstracted and run on multiple hardware platforms, including digital computers" (Mateas 14). If one develops this idea further, human brains are not seen as the only home to higher intelligence, but machines could function and think just as humans do. Interactionist AI, in contrast, stresses the need of a body to perform

certain intelligent actions. Thus, it does not focus on the 'intelligence' needed for a certain action, but views the AI creature's body as an integral agent interacting in a situation and therefore defining its intelligence (Mateas 10). Hence, an AI creature's body is "necessary even for forming abstract concepts; abstractions are based on sensory-motor experience" (Mateas 10). It follows that the body-as-agent forms the mind's structure – if the body changes, the mind does too (Mateas 11).

The definition given above focuses on problem-solving. However, there must be more to human intelligence than mere problem solving abilities before any AI could be considered indistinguishable from human intelligence. *A Dictionary of Psychology* gives another definition of AI as:

> The design of hypothetical or actual computer programs or machines to do things normally done by minds, such as playing chess, thinking logically, writing poetry, composing music, or analysing chemical substances. The most challenging problems arise in attempting to simulate functions of intelligence that are largely unconscious, such as those involved in vision and language […]

This definition includes more aspects than the one given above, as it describes intelligence as something more than pure intellect. The writing of poetry and composition of music surpass the AI definition and abilities already discussed. Furthermore, the unconscious level of the mind is said to play a part in AI as well, involving vision and language. The implication of this definition shall be discussed in depth at a later point; it is crucial to note, however, that when adding it to the more classical one given above, AI creatures would have to have some unconscious aspect to their intelligence in order to classify as artificial intelligent. Those are harder to control and understand than conscious decisions one makes based on pure intellect. If one wants to describe the creatures in the literature investigated here as artificially intelligent according to the definition above, they would necessarily also have unconscious functions of intelligence and emotions.

If AI also includes this definition of intelligence on an unconscious level, the question arises whether AI can be clearly distinguished from human intelligence. Alan Turing's famous test describes this specific difficulty and offers a way of clearly defining AI. Mateas (7-8) describes the procedure of said test as follows:

> In the Turing Test, a human judge engages in typed conversation, through a terminal, with both a human and a machine that are present in another room. The judge must determine, based on the responses to her typed queries, which is the human and which is the machine. If the judge can't tell the difference, we deem the machine 'intelligent'.

The Turing Test posits that an advanced AI being's answers to typed questions would be indistinguishable from a human's. Considering this, the question arises; what characterises a human as an intelligent being?

2.2 Human intelligence - what makes us human?

2.2.1 Different definitions

Human intelligence can be defined with various different approaches and remains a much discussed field of research. *A Dictionary of Psychology* characterises intelligence as "cognitive ability". A brief history shows the varying definitions of human intelligence:

> 'the ability to carry on abstract thinking' (Lewis Madison Terman, 1877–1956); 'the power of good responses from the point of view of truth or fact' (Edward Lee Thorndike, 1874–1949); and 'the capacity to inhibit an instinctive adjustment, the capacity to redefine the inhibited instinctive adjustment in the light of imaginally experienced trial and error, and the volitional capacity to realize the modified instinctive adjustment into overt behaviour to the advantage of the individual as a social animal' (Louis Leon Thurstone, 1887–1955). Since then, one of the most

> influential definitions was the one put forward in 1944 by the Romanian-born US psychologist David Wechsler (1896–1981): 'the aggregate or global capacity of the individual to act purposefully, to think rationally, and to deal effectively with his environment.' (*A Dictionary of Psychology*)

Despite using different formulations and focusing on various aspects of intelligence, these definitions can all be said to involve intelligence on a level of problem-solving and rational thinking. Friedenberg (109) observes that "an intelligent agent" must be flexible in a changing environment, regardless whether the agent is a machine or a human being. When thinking back to the description of artificial intelligence the definitions given here do not seem to differ strikingly from AI. If artificial intelligence is meant to mimic human intelligence, it has to incorporate all these aspects, and they should be detectable in an artificially intelligent creature.

Another aspect of intelligence is emotional intelligence, which is sometimes treated as a specific sub-category of intelligence. It is defined as follows: "[the] [a]bility to monitor one's own and other people's emotions, to discriminate between different emotions and label them appropriately, and to use emotional information to guide thinking and behaviour" (*A Dictionary of Psychology*). Goleman (34) formulates this concept somewhat more clearly, when he states that being emotionally intelligent entails being capable of controlling one's emotions and preventing them from overwhelming one's thoughts. Emotional intelligence thus acknowledges the importance of emotions to influence thoughts and behaviour. Goleman (8) explains that humans have two minds, one being emotional and the other rational. While the rational mind is typically concerned with conscious processes, the emotional mind seems to be rather impulsive. This ties in with the second AI definition given in the previous section, which mentions an unconscious part of intelligence. Emotions appear on both a conscious and unconscious level of the mind. However, if unconscious emotions influence one's actions, this can lead to problems, which shall be discussed later. It can be said here, however, that if one

aims at comparing AI to human intelligence, one must also consider emotional intelligence and the unconscious actions it entails.

2.2.2 Characteristics of human nature

Considering the various definitions of intelligence, one might ask, whether there is not something more to being human than having outstanding intellectual abilities. Dennett ("Conditions of personhood" 177-178) introduces a set of six criteria which constitute what it means to be human. They are as follows:

1. Persons are rational beings…
2. Persons are beings to which states of consciousness are attributed, or to which psychological or mental states are ascribed.
3. A person is a being to which an attitude or stance of personhood is taken …
4. … to be a person is to treat others as persons, perhaps in a moral way.
5. Persons must be capable of verbal communication or language. …
6. Persons are conscious in some special way that other species are not.

These criteria go beyond the scope of definitions of intelligence. They show that in addition to the rationality of the human race, other factors have to be fulfilled in order to be classified as human beings. Thus, one can ask whether an AI creature only needs to fulfil the criterion of rationality in order to be comparable to human beings, or whether it must, in fact, match all the other aspects mentioned. Friedenberg (5) defines an artificial person as being indistinguishable from a human being with regards to its actions and behaviour, although its outer appearance might differ. Furthermore, he states that, if tested, an artificial person's behaviour would be exactly the same as a human's (5). Considering this, it seems that AI creatures do have to fulfil all the given criteria of humanity.

Thus, the notion of artificial intelligence alone might not be enough to analyse an artificial creature's similarity to humans. Therefore, the above list of criteria shall be of great

importance in the later chapters of this thesis.

3 AI and human intelligence applied to the protagonist

3.1 The novel's realities

Before one can turn to a deeper discussion of the individual characters described in ASO and IR the setting of the novels shall be described briefly. They were written in the 1960s, and are both set in an imaginary future. Although ASO begins with the very dawn of mankind, the timespan this thesis is primarily concerned with lies in a fictional future. It starts with the time HAL was built in 1997 (ASO 202) and spans to his actions on board the spaceship *Discovery*. However, as classified by Seed, one should not see science fiction literature as being a forecast of the future (Seed 1), but rather as "… an embodied thought experiment whereby aspects of our familiar reality are transformed or suspended" (Seed 2). Therefore the novel's plots will not be discussed as if they were describing a possible future, but rather, taken as said thought experiment; a parallel reality. In ASO's reality, mankind is able to travel into space, build strongholds on the Moon, and take a long journey to Saturn. Hence it is also fitting in this reality to assume that computers are far more advanced than the ones we are used to in our reality. HAL is described as "a masterpiece of the third computer breakthrough" (ASO 116) and as such is the most advanced computer on the market in the novel's fictional reality.

IR's reality spans from the year 2008, when Susan Calvin joins U.S. Robots, to the year 2052 when Stephen Byerley is world-coordinator in his second term. Throughout this period of time, the development of robots is described, beginning with the somewhat simplistic Robbie and ending with the Machines, running a significantly changed world. Although Robbie might seem cruder than the robots at the very end of the novel, he is a robot of a technical status which surpasses any robot of our reality. What is more, IR's technical

standards as a whole surpass ours by far, as space-travel and extra-terrestrial use of robots is described and seen as ordinary. Therefore, both novels describe an advanced reality, in which artificial intelligence exists and is a part of everyday life.

Taking this fact into account, one cannot analyse the novel's characters according to the standards of current AI research. Friedenberg, for example, analyses that there is no AI in our world which surpasses the human being in all respects, but that it might be possible to create an artificial being if the technological development advances and becomes more complex (Friedenberg 242). A similar possible development is mentioned by Kuck (499) when he explains that there have been enormous advances in computer sciences since ASO's release, but that a HAL has not yet been built. In contrast to our technological world, in the novel's realities, these artificially intelligent beings do exist and are worked with. Therefore, since their existence is taken as given in the reality of the described novels, it will not be necessary to speculate about how such creatures could be built or function in this thesis. HAL and the robot's technical realisation will be taken as granted, it is their actions and psyche which shall be analysed and discussed. To begin with, however, their status as AI creatures shall be described in the next section.

3.2 HAL

HAL, the on-board computer serving the spaceship *Discovery* on its voyage as told in ASO can clearly be shown to be artificially intelligent. To begin with, HAL would pass the Turing test, which is also mentioned in ASO (ASO 119) itself. Furthermore, when describing HAL, Stork explains that "the overarching technical issue associated with HAL […] is his artificial intelligence (AI)" (5). He is a computer carrying out the routine work of a whole spaceship whilst monitoring and caring for the hybernated crewmembers (ASO 118). Considering the two definitions of AI given above, HAL definitely is "self-aware" and "sentient" and also has unconscious parts of intelligence, which cause him to question his obedience and to take

action himself, namely killing crewmembers.

If one wants to classify which type of AI HAL represents, one should consider that he does not have a body like a robot which he could use to interact with the environment. Nevertheless, the ship could be described as his body, as he can control it in various ways. However, HAL does not need the Discovery in order to think. Mateas (11) thus classifies him as an example of classical artificial intelligence because his intelligent actions stem from an internal process. HAL, therefore, can be classified as an example of classical AI.

3.3 The Robots

In IR, there are nine different robots to consider (namely: Robbie, Speedy, QT-1, Dave, Herbie, Nestor 10, The Brain, Stephen Byerley and The Machines). In order to classify them as artificially intelligent, one can, firstly, apply the Turing Test again. All robots would pass this test easily and are, except for Robbie, equipped with the ability to speak and even have human-like voices which might make passing the test easier for them; they also are described to have an intelligence, which, in case of the later robots, succeeds that of humans.

Robbie is one of the first robots and is not able to speak. However, Robbie does understand speech and shows a variety of emotions throughout his story. He feels sad when Gloria accuses him of cheating in a game of hide-and-seek (Asimov 7), handles Gloria carefully (Asimov 8), which lets one assume that he cares for her and nods his head with excitement (Asimov 9) when Gloria tells him a story. This array of emotions classifies him as AI. Furthermore, he is designed to be a robot-caretaker for children, playing with them but also protecting them, for which he requires intelligence. Robbie, thus, certainly can be said to be an AI creature.

Speedy is a highly specialised robot, designed to work on Mercury to quarry selenium so that humans can survive in the hostile atmosphere of the planet. In order to do this, he has to be intelligent. Furthermore, he gets into somewhat of a crisis, when he has to decide

whether to put himself in danger or to fulfil his orders. This alone shows that Speedy must be of a higher form of intelligence, as a machine without intelligence would not question an order.

QT-1, then, is used on a space station and develops transcendence, as he questions where he came from and who made him. He reasons that humans cannot have created him as they are far less intelligent. His conclusion, that there must be a God-like master shows higher intelligence and superior thinking skills.

Those higher thinking skills are also used by Dave, who works in a mining station and has six robots below him, who he is in charge of. When these robots are endangered, he develops a certain habit to keep them from harm, which shows that he is capable of developing a psychological response to threatening situations. This happens on an unconscious level, as Dave is not aware what he is doing, thus proving that he has some part of unconscious intelligence level.

Herbie, the mind-reading computer, is defeated in the end by his own empathy and incapability to hurt humans, forcing him to lie. Empathy and lying are both signs of intelligence, making Herbie almost indistinguishable from humans if it weren't for his appearance.

Another clear sign of intelligence is portrayed with Nestor 10, who, similar to QT-1, questions whether he has to serve a master inferior to himself. He even decides to harm this master and tries to outthink the humans.

The Brain is capable of overseeing the construction of a whole spaceship using a technology never applied before. This alone shows its superior intelligence. In addition to that, it develops a sense of humour, which is another sign of emotional intelligence.

Stephen Byerley, then, is indistinguishable from a human being. He looks and acts like one, has a human job, and even is World-Coordinator in the end. Therefore he clearly is an example of AI.

Finally, the Machines running the world at the end of the novel surpass human knowledge by far, and decide what is best for the entire human race, which makes them clearly artificially intelligent.

Summing up, all the robots in IR can be classified as AI. Thinking back to the two types of AI described above, most of robots can be described as interactionist AI, as they clearly are "embodied agents interacting in a physical or virtual world" (Mateas 10). Mateas (10) claims that if the bodies of interactionist AI creatures change, so does their intelligence. As IR describes the history of robotics, the creatures mentioned in it, improve throughout the story. While Robbie still seems somewhat cumbersome and his sole task is to care for Gloria, Stephen Byerley is indistinguishable from a human, and aims to win a human election. Thus, one can say that as their bodies improve, their minds improve too.

3.4 The blurry boundary between AI and human intelligence

After having defined artificial and human intelligence, the question arises whether these intelligences can be clearly distinguished from one another or whether they are, in fact, closely linked. Friedenberg (9) mentions, that when thinking of AI, one assumes a significant difference from human life, however, from a scientific viewpoint there should be no distinction between the two. If one regards the descriptions of the artificially intelligent characters in ASO and IR, the texts seem to struggle to define a clear boundary to human intelligence. HAL is first described as follows: "[…] a machine intelligence that could reproduce – some philosophers still preferred to use the word 'mimic'- most of the activities of the human brain, and with far greater speed and reliability" (ASO 117). However, HAL's programming is not described as such but ASO (117 – 118) uses the vocabulary of a human in school or in training, saying that he is "trained" instead of programmed, that the spacemen are his "colleagues" rather than commanders, and even referring to his "electronic childhood" rather than the production process of a machine. These descriptions depict HAL's intelligence

as something learned rather than something programmed, making him seem more human than machine. Hence it is impossible to classify and analyse him as if he were a mere machine.

When looking at IR, the robots evolve from rather simple thinking agents into more complex characters, whose thought processes are very much comparable to those of humans (Majed 1). Throughout the novel, the humans working with them are dealing with various problems their advanced brains cause. Susan Calvin is described as "robopsychologist" (IR) in the novel. This term shows that robots are assumed to have a psyche, which can develop disorders very similar to those of humans. Thus the boundary from artificial intelligence to human intelligence is blurred again. Majed (2) even goes as far as saying, "They [robots] demonstrate the human qualities of minds and soul, reason and emotions", thereby indicating that robots do share these qualities. This means that in the novels discussed, both the Robots and HAL surpass the level of being a mere machine and evolve into something more.

The ability to feel emotions and to be able to control them seems to be a uniquely human trait, as their brains are structured in a certain way (Friedenberg 194). Thus, if AI creatures were capable of having humanoid emotions, their brains would have to be structured similarly. To illustrate the connection and importance of emotion in human intelligence, Picard (295) talks of a patient having a type of brain damage where the connection of the limbic system and the neocortex, which plays a key role in feeling emotions, is blocked. This causes the patient to seem unemotional and purely rational. One might, therefore, assume that this patient would have no difficulty in making rational decisions, since he or she is not hindered by emotions. However, Picard (295) contradicts this prediction, as the patient is described to have severe problems in decision-making processes because he or she cannot link bad experiences and feelings with situations, nor connect positive feelings with a decision. The patient could thus be described as too rational; he or she can only think rationally but is not capable of letting emotions influence his or her decisions. This 'over-rationality' can also be found in AI creatures, as they are commonly not assumed to have emotions. Therefore they

can be experts in a certain field of knowledge but struggle with making decisions (Picard 296).

However, if one considers the protagonists of the two novels discussed here, they do not seem to have difficulties to reach decisions, and their decisions are even influenced by emotions. HAL decides to kill his human colleagues, while being heavily influenced by emotions. What is more, Dave asks himself if HAL can feel pain, when he starts to pull out his brain circuits (Clarke 201). This reveals that he thinks him capable of feeling emotions. The Robots also display the ability of making decisions which are influenced by emotion, for example when Q-T locks away his bosses when he feels they disrespect the master or, when Nestor 10 feels superior and decides to hide from his bosses. This shows that the creature's intelligence cannot simply be defined as being artificial, but that the boundary between human and artificial intelligence is blurred. It follows that their actions cannot only be analysed with an AI research background, but can also be studied with the human psyche in mind.

3.5 How to analyse the novel's protagonists

Taking all the insights of the former chapters into account, the protagonists of ASO and IR can be analysed and compared to each other. As mentioned, HAL classifies as classical AI, while the robots represent interactionist AI. These two forms of AI cannot be clearly separated from each other but interactionist AI can be seen as evolving out of classical AI. One could imagine a timeline of AI evolution, and place ASO's HAL before IR's Robots on the line. Regarding this evolution, Friedenberg (115) mentions that as intelligence grows, a shift happens where the being's actions are influenced more by internal processes than by its surroundings. These internal processes could be interpreted as unconscious processes or higher levels of intelligence resembling human intelligence. The mentioned timeline, thus, is useful to analyse this thesis' protagonists.

Therefore HAL's actions will be thematised first, and discussed under the aspects of

somewhat more basic human markers of intelligence. Then the robot's stories will be analysed and compared to some more advanced features of intelligence.

4 *2001: A Space Odyssey's* HAL

To analyse whether HAL's intelligence really resembles a human form of intelligence, three markers of intelligence will be discussed and applied to him in the following sections: planning, lying, and having emotions. HAL's description in the novel paints him as outstandingly intelligent, but his human colleagues sometimes seem to be uncertain whether to see and treat him as a machine or a living being. This becomes clear in the novel when Dave and Frank see HAL as their colleague and feel embarrassed to talk about him while he listens, (ASO 170) and again when Dave muses about HAL's ability to feel pain when he removes his memory blocks in the process of switching him off (ASO 201). It is HAL's advanced intelligent behaviour which causes this blur.

4.1 Ability to plan – common sense and predictions

Wilkins (309) argues that in order for creatures to survive in a certain environment, fitting behaviour is essential. This calls for well thought through plans. Planning, thus, constitutes a part of intelligence, as it involves thinking ahead and designing an action for a future which has not yet happened (Friedenberg 122). Hawkins and Blakeslee (65) go even further and claim that the brain's ability to predict what will happen, depending on its previous experiences and on the given situation, is the "foundation of intelligence" (Hawkins and Blakeslee 60). In accordance with Hawkins and Blakeslee, Friedenberg states that the ability to plan ahead is unique to the human species and has secured its survival (122).

At first planning might seem to be a very simple process, however, Wilkins reveals that it involves many features of intelligence. "Intelligent action requires us to reason about the future consequences of actions and form plans to achieve goals. This ability develops in

childhood and is part of what we call common sense." (Wilkins 309). This common sense is defined by Friedenberg (27) as a certain understanding of the world which is regarded as self-evident. This is what makes planning difficult. Humans are not born with common sense, but acquire it throughout their lifetime. The problem is, that plans must often be made in the presence of "conflicting desires" (Wilkins 309). For example, if there is a fire in a high building and one wants to escape as quickly as possible, one might be tempted to use the elevator instead of the stairs assuming that it might be quicker. However, taking the elevator might cause an additional life-threatening situation, which is why one hurries down the stairs instead. This sounds simple but actually involves the common sense mentioned above. Wilkins (309) also states that "Plans must often be based on incomplete information, so the planner has to make reasonable assumptions about the current state of the world and the consequences of future action". In the fire-example from above, one does not know whether the fire is really life threatening and that it is not safe to use the elevator. One must, therefore, assume what usually happens in a house-fire, and what the best action in such a scenario is. The "central problem", according to Wilkins (309), "is predicting and evaluating what will result from executing a given plan". Again, in order to do this, common sense is needed. To sum up, one can say that planning is a complicated and fragile process, even for human beings.

In ASO there are, first of all, human plans. The Discovery is on its way to Saturn on a mission planned by human beings. Its human crewmembers operate on a very specific plan every day. The mission controllers on earth certainly think that they have taken care of every possible thing, which could go wrong with the mission and, as Wilkins (314) says, written programmes for every probable problem. HAL himself does not need the ability to plan for his every day duties (Wilkins 314). However, this does not hold true as HAL begins to have his very own individual plans. So the flaw in the otherwise maybe perfectly planned Saturn mission was the on-board computer's mind (which should have executed all the plans).

Planning can also be described from an AI research perspective. Wilkins (311) describes the reflection of actions and their impact on the future in order to "synthesise" a plan as the core problem of artificial intelligence planning. He says that synthesis is difficult, as there can be many ways to approach a problem, with different consequences. To give a brief example of this difficulty, Wilkins (314) describes how a computer approaches a problem: "Before a computer can solve a problem, it must be able to represent it". An example from AI research is the Sussman anomaly, shown in figure 1.

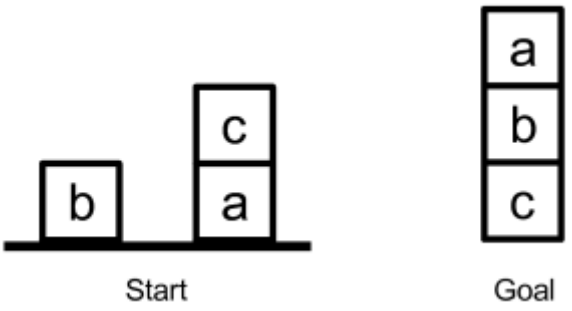

Fig. 1 Sussman anomaly from Wernhard, Christopher. "Liner Planning System - User Manual (Draft); infraengine.com; Feb, 5 2016, *http://www.infraengine.com/doc/planner/planner_manual.html*

Wilkins describes it as follows:

> The problem is an anomaly because you cannot do either of the necessary actions first without having to undo it later. If you put A on B (after putting C on the table), you will have to take it off again so that you can move B onto C. If you put B on C first, you will have to undo this to move A. This problem shows us that it is not sufficient to achieve goals in some particular order; the best plan requires the planner to interleave the actions to achieve the two goals. (Wilkins 315)

The solution to the Sussman anomaly may seem logical to humans, for computers, however, the movement of blocks causes problems as they do not know that "one block cannot be moved on top of itself"(Wilkins 315). This must be told to the computer explicitly. Humans would know this because they are able to sequence actions in order to solve problems (Friedenberg 197). Another fact which computers normally do not know is that objects such as blocks do not change their location once they are put down somewhere (Wilkins 316). For

example, if one would move block C first, a computer would not automatically assume that this block stays were it was put down until it is moved again.

Another AI research problem describing computer's lack of common sense is the Yale shooting problem. In this problem, described by Wilkins, one assumes that one agent, in this case HAL, has a loaded gun. Opposite to him is another agent, in this case Dave. The only one who can act is HAL and he knows two facts "Dave is alive" and "the gun is loaded" (Wilkins 319). HAL now has two options: he can either wait or shoot. Waiting would cause no change. Shooting with a loaded gun would result in Dave's death. Computers do not know that Dave will die after the action 'wait' followed by the action 'shoot' (Wilkins 319). Wilkins (319) explains this saying, "The problem for computers is in logically specifying the notions that things typically stay the same unless there is information to the contrary". Humans would immediately know that nothing happens unless HAL fires. A computer, in contrast, might come to the explanation that the gun could unload automatically during 'wait' (Wilkins 319). This would happen because of the lack of world knowledge.

Considering ASO, the Sussman anomaly and the Yale shooting problem demonstrate that HAL cannot be seen as a computer which cannot think into the future, but must have some common sense similar to humans. His actions in the novel cannot be explained as being programmed by mission control, so he must be planning on his own. Stork (9) lists HAL's plans: "to test whether the AE35 unit was in fact faulty, to navigate the ship … to kill the crew". As the AE 35 unit was not actually broken, HAL could have used its supposed malfunction as a first step to kill the human crewmembers. In order to do this, he would have had to know that if he predicted a malfunction, Frank would leave the spaceship to replace the unit. Furthermore, he must have reasoned that as Frank leaves the spaceship, he is vulnerable and thus easier to attack. He also must have known that the hibernated crewmembers would die without oxygen when he opened the spaceship's airlocks. This consideration of the causes certain actions will have in the future proves that HAL must have some kind of common

sense, or world knowledge. Norman (265) argues that HAL must possess the capacity to understand human behaviour.

HAL's ability to use common sense shows that he is able to learn and does not rely solely on what he has been programmed to do. How HAL could have acquired this knowledge and learned to reason can be explained with Hawkins' and Blakeslee's theory about human intelligence. They describe the neocortex of the human brain as being an "organ of prediction" (Hawkins and Blakeslee 60). To undermine this theory, an evolution of the human brain, from a reptile's brain without a cortex, to the human brain we have now is outlined (Hawkins and Blakeslee 61-72). The neocortex of humans is linked to the sensory system of the brain, and can therefore store memories of specific situations and how to react to them (Hawkins and Blakeslee 67). The brain can then make predictions about what will happen in a similar situation. What is unique to the human physique of the brain is that the neocortex influences most of our body's motoric system and therefore plays a key-role in our behaviour (Hawkins and Blakeslee 70). Hawkins and Blakeslee, consequently, claim that the ability to make predictions is the key to human intelligence as "[t]o know something means that you can make predictions about it" (70).

Considering that HAL is a computer, without a human brain, one might assume that Hawkins's theory cannot apply. However, HAL's nervous system is described as a brain in the novel (ASO 201). When Dave decides to switch off HAL, a separation between "higher centers" and "purely automatic regulating systems" is described (ASO 201). This distinction is similar to the human brain's complex neocortex and its' somewhat simpler motor system. Finally, Dave even compares himself to a brain surgeon operating on the human cortex (ASO 201). Thus, HAL's ability to predict future events and plan ahead can be explained by the complexity of the brain's neocortex. Furthermore, HAL's reasoning with the help of common sense, or world knowledge could also stem from the learning processes of his brain's memory.

HAL's plans are flawed, as well as those of his human colleagues. Norman (265) states that creativity is needed to alter plans in which something unexpected happens. In order to be creative, one has to rely on past knowledge and then combine the 'already known' in new ways (Friedenberg 129). Hawkins and Blakeslee (124) interpret creativity as "making predictions by analogy", and define it as something humans use on a daily basis without noticing it. They postulate, similarly to Friedenberg, that creativity is ordering learned patterns in a new way in order to adapt to a situation (Hawkins and Blakeslee 126). After HAL kills Frank he does not expect Dave to survive the opening of the airlocks, and is consequently switched off by him. One could say that Dave was very creative when he managed to survive in a spaceship without oxygen. He had to use his world knowledge in order to quickly find the emergency-room with oxygen. However, HAL also exercised creativity by opening the airlocks in the first place. His action is a quick and creative way of disobeying Dave's command for control over the spaceship. HAL must have reasoned, that Dave's reaction to an opening of the airlocks would guarantee him longer control over the ship. This also proves HAL's ability to learn from past experiences and to create the analogies Hawkins and Blakeslee refer to.

To summarise, HAL's planning "stems from a general intelligence and an ability to learn" (Wilkins 330) and his brain structure allows him to make predictions about future events, enabling him to plan according to them. Mission control did certainly not programme him to kill his human crewmembers, but he used his intelligence to plan actions which they did not predict (Wilkins 330). HAL, thus, has the ability to plan and consequently fulfils this marker of human intelligence.

4.2 Ability to lie – social inference and schemata

Humans are the only known species which are able to lie, apart from some monkeys, which can only do it partially (Norman 266). According to Norman (265), "Lying is at the pinnacle

of human intelligence, because it requires not just knowledge, but meta-knowledge". This meta-knowledge means that if one tells a lie, he or she has to know what the person lied to is going to find plausible, how the story might fit into the context, and how it could be altered spontaneously under certain conditions (Norman 265). In order to consider this, a deep understanding of the human mind and feelings is needed and, according to Norman (264): a "shared cultural background". These abilities can be referred to as social knowledge. With social knowledge one is able to carry out social analysis in order to analyse and interpret the feelings and behaviour of others (Goleman 118). This understanding does not develop overnight, but is acquired through learning and development. Norman (265) mentions the education system as having a great impact on the knowledge of social conventions of humans.

ASO's HAL has not been to school, and thus has not learned social conventions. However, he is programmed with a vast knowledge of humans. The programming shows in his chess play which is described as follows: "If HAL went all out, he could win any one of them [games]; but that would be bad for morale. So he had been programmed to win only fifty percent of the time, and his human partners pretended not to know this" (ASO 127). HAL's knowledge of how to play human games, ties in with the awareness of another aspect of social knowledge, namely, social schemata. These are defined as connected cognitions which enable humans to quickly come to an understanding of a certain person or situation (Hogg and Vaughan 49). When HAL plays chess he could therefore activate a schema of chess-playing and know how to behave in this specific setting. The schema would also tell him that winning every game is bad for morale.

One could argue against HAL's chess playing as an example of his social knowledge and see it simply as an example of excellent programming. The programming of schemata could be possible in ASO's reality and would therefore not mark HAL as being intelligent. However, he is able to take his programmed abilities and develop them further or even use them against humans. This can be seen in the scene after Frank's death, when Dave requests

to have manual hibernation control from HAL. HAL argues against this by saying “I can tell by your voice harmonics, Dave, that you’re badly upset. Why don’t you take a stress pill and get some rest?” (ASO 187). Dave then still insists on having full control, which HAL dismisses by saying “I’m sorry Dave, but in accordance with special subroutine C1435-dash-4, quote, When the crew are dead or incapacitated, the on-board computer must assume control, unquote. I must, therefore, overrule your authority, since you are not in any condition to exercise intelligently” (ASO 187). In this example HAL is applying a social schema to a situation, namely what humans normally do if they are in pain or distress. HAL not only uses a social schema, but also manipulates it to serve his purposes by interpreting Dave’s distress as an excuse to take over the spaceship. This shows that HAL is able to make social inferences, which are defined as using inferential processes to judge a certain situation based on the social information one has (Hogg and Vaughan 68).

Thus, HAL has some of the described social knowledge, as he can interpret emotions correctly. Furthermore, HAL can be said to take this knowledge and use it for his own purposes. When he quotes the ship’s emergency guidelines to Dave, he uses his crewmate’s state of mind to argue for his seizing full control. HAL is therefore shown as able to manipulate a setting with the help of social knowledge, which is exactly what humans do when they lie.

A first example for a deliberate lie is HAL assuring Dave that he is under full hibernation control after the argument described above. As Dave does not give in to HAL’s suggestions to take a stress pill or to leave control to him, HAL seemingly surrenders when he says, “O.K. Dave … You’re certainly the boss. I was only trying to do what I thought best. Naturally, I will follow all your orders. You now have full hibernation control” (ASO 187). At first HAL seems to tell the truth as Dave starts to awaken his crewmembers. But HAL’s lie becomes obvious when he opens the Discovery’s airlocks and thereby kills all the hibernators and attempts to kill Dave. He has not followed all of Dave’s orders but has deliberately

deceived him into thinking Dave had the upper hand. HAL must have known that Dave would immediately start to awaken the hibernators, which would give HAL the chance to open the airlocks unnoticed until it was too late to stop the opening process. Consequently, HAL's lie about the hibernation control is an example, which proves his ability to estimate how a human being might react to a lie and to use this knowledge to his advantage with the help of social inference.

Another example for HAL's ability to lie is the malfunction of the AE-35 unit for the second time. After the astronauts have already exchanged the unit once, HAL predicts its failure again, following a conversation by Frank and Dave about maybe switching to earth control due to HAL's error. HAL tries to convince the astronauts of the faulty unit twice. The first announcement is rather unemotional, he simply states, "We have another bad AE-35 unit. My fault predictor indicates failure within twenty-four hours" (ASO 171). The astronauts react sceptically and report to mission control, who confirm that HAL has some problem and therefore suggest switching to earth control. As mission control starts to explain the steps to switch HAL off, he interrupts the message by saying, "Condition Yellow! Condition Yellow!" and "The AE-35 unit has failed, as I predicted" (ASO 175). HAL could be lying about the second failure because he overheard the astronauts' musing about switching him off. Dennett ("Computer ethics" 363) mentions that HAL does not know that he can be switched on again, once he is switched off, and therefore assumes a shut-down to mean his death. HAL could thus have decided to lie to 'stay alive'.

He could have assumed that Frank would leave the spaceship again to repair the unit if he lied that it is broken, just as he did with the first unit. Maybe HAL planned this as the perfect opportunity to kill Frank. However, his first announcement does not have the wanted effect, as the astronauts do not believe him immediately, but double-check with mission control, who tell them that the problem apparently is not the AE-35 unit, but HAL himself. As HAL hears that, they begin to give immediate instructions to switch him off, he reacts by

stating that the unit has now failed completely. Despite mission control's pledging for the unit's functioning, Dave and Frank believe HAL when he shows them a picture which, "for the first time since the beginning of the voyage … had changed. Earth had begun to drift from the cross-wires; the radio antenna was no longer pointing toward its target" (ASO 175). After seeing this, the astronauts do not question HAL but trust him, and Frank ventures out to repair the unit once again. HAL thus succeeds to convince them with a lie. The novel does not explain how HAL has managed to show a wrong picture of the antenna. He could have manipulated the picture or he could have interrupted the connection in some way. Whatever he did, he lied. He must have understood that his first attempt was too feeble, as it only inspired the astronauts to contact mission control, so he reacted by faking an emergency. HAL could have assumed that Frank and Dave would delay his shut-down if he gave them an emergency signal, as they would need to repair the connection to earth in order to save their lives. Therefore, HAL used his social knowledge to predict how the humans would react to his lie.

Furthermore, mission control mentions something very interesting. They state, "We have completed the analysis of your AE-35 difficulty, and both our HAL Nine Thousands are in agreement. … As we suspected, the fault does *not* lie in the AE-35 unit, and there is no need to replace it again" (ASO 174). This shows that the other HALs on earth were used to check on the HAL on board the Discovery, and they worked fine. Thus HAL's prediction about the failure of the unit must be a lie. If these computers had the exact same programming as HAL, HAL must have evolved and learned to use his social knowledge in order to be able to lie. Consequently, something must have motivated HAL to act differently than his twins on earth.

The given examples show that HAL has the ability to lie and that he uses his knowledge of humans to his advantage. He is even able to convince them when they are informed about his malfunction by other humans. Thus HAL passes the second marker of

human intelligence discussed here.

4.3 Ability to feel – HAL as a conscious being

The last marker of human intelligence discussed here is the ability to have emotions. Human behaviour is influenced by emotions most of the time, as they play a central role in our decision-making processes (Goleman 4). Decisions are often made based on how one 'feels about something', whether one trusts or distrusts another person is decided by good or bad feelings towards this person, and so on. Goleman (4) even goes as far as saying that the term 'homo sapiens' is misleading because it paints humans as purely rational beings, when they are, in fact, heavily influenced by emotions. Computers, in comparison, normally do not have feelings and instead, express their state of mind per codes (Picard 280). The question now is whether HAL, being a computer, has feelings and if they also influence his decisions.

To begin with, Picard (281) states that it is difficult to define the term emotion because of its range. Does it include human states of mind such as fear and anger, which are termed affective states, and contain things such as love? Picard (281) simplifies by using the terms "emotional computing" and "affective computing". She differentiates between the two by linking emotional computing with a rather negative connotation, meaning a reduced ability to act rationally because of emotions; and describing affective computing as taking emotions and influencing them deliberately (Picard 281). One could hypothesise that humans are both affective and emotional, depending on the situation. ASO's HAL could be programmed to only be affective – he should seem like a human crewmember but still remain able to act rationally.

HAL is clearly able to interpret the emotions of his fellow crewmembers. He works together closely with Frank and Dave, sees them interact and is therefore familiar with their

behavioural patterns (Picard 283). However the question arises whether HAL himself acts emotionally or according to his programmed affective computing.

In "Between Planets" (part three of IR) HAL is not particularly noticeable during the description of *the Discovery's* voyage. He fits into the routine on board the spaceship and only speaks very little. He watches over everything, occasionally engaging with the crew in a game of chess and they seem to trust him completely. Even when the Discovery passes Jupiter, HAL only speaks to help and guide the crew. His utterances in this part of the book sound rather technical, such as "Earth signal is fading rapidly. … We are entering the first diffraction zone" (ASO 141). One could describe this part of ASO as featuring a fully functional HAL, who watches over the spaceship and its crew. However, the situation changes in the novel's fourth part, "Abyss". HAL begins to predict the malfunction of the AE-35 unit. The first announcement of this malfunction is still similar to HAL's utterances in "Between Planets". He declares, "I am having difficulty in maintaining contact with Earth. The trouble is in the AE-35 unit. My Fault Prediction Center reports that it may fail within seventy-two hours" (ASO 150). HAL gives a clear technical description of what he detects and the astronauts immediately trust his judgement. When they ask him what procedure he would suggest, HAL proposes to exchange the unit, which means leaving the ship. Before doing so, Frank and Dave check with Mission Control, who give them the permission to operate, also trusting HAL's abilities. The trouble on board begins, when the astronauts find out that the unit was not broken after all, but that HAL must have made a mistake. This seems curious as HAL is built to be an error-free computer. Nevertheless, just like the famously unsinkable Titanic, the computer incapable of mistakes happens to make mistakes.

Why these mistakes happen can be explained with the help of the psychology of human emotions. To begin with, they could be caused by "chronic repetitive worrying" (Goleman 65) stemming from clashing loyalties. HAL knows the Discovery's true mission while the two awake crewmembers, do not. On the one hand, he has been programmed to carry out this

mission, expresses his loyalty and enthusiasm for it several times, and aims to fulfil it to the best of his abilities. On the other hand, HAL is loyal to his crewmembers. This situation could cause HAL to worry endlessly about what to do and to whom to be loyal. A possible explanation for HAL's mistake could therefore be that "… HAL experiences an internal state of conflict about the mission, conflict between the real mission and being forced to conceal it from Dave and Frank. HAL broods over his predicament until he begins to make errors" (Picard 300). This constant state of brooding or worrying about something is described by Goleman (65) as "chronic worry", a situation in which, similar to Picard's description, the mind fixates on the problem while blocking out everything else. The human crewmembers were not supposed to be told about the mission's true purpose until they had almost reached Saturn to prevent any such psychological disorders. HAL was apparently trusted to handle this situation easily, but, considering his actions, he failed to do so. Picard explains that HAL's inner conflict could take up so much of his 'computational space' that errors happen: "If most his resources are allocated to protection and trying to reason about a source of distress, less capacity for diagnosing and treating the operations of the ship will be available, which increases the probability of error" (Picard 300). This means that HAL could have been silently worrying about his inner conflict throughout the first part of the Discovery's journey until he had put so much thought to his own problem, that he could no longer successfully fulfil his duties. HAL's distress could be even worse as he knows that he is a computer incapable of making mistakes, and therefore could not admit his false prediction about the AE unit, even if he would have noticed it to be a mistake. Consequently, HAL decides to propose to exchange the unit so that he can cover up his error.

When Frank and Dave test the AE-35 unit, HAL has said to be faulty, they find out that he must have made a mistake, to which mission control also agrees and suggests that they might switch control over to the computers on earth. They state that "… we may have to disconnect your nine-triple-zero temporarily for program analysis" (ASO 168). As HAL can

hear everything which is said on board, he also receives this message. While HAL does not comment on it, the astronauts discuss the matter but speak in hushed voices and try to not talk openly about a switch to earth control. The reason for this is given in the novel, “HAL was their colleague, and they did not wish to embarrass him” (ASO 170). The fact alone that the astronauts hold HAL capable of experiencing an emotion like embarrassment is proof that they assume HAL to have emotions in general. However, they clearly did not think that this would have any further effect. Despite the careful handling of the possible switch off, HAL interprets it as a threat (Picard 301). His thoughts are described as follows. “He had been threatened with disconnection; he would be deprived of all his inputs, and thrown into an unimaginable state of unconsciousness. To HAL, this was the equivalence of Death. For he had never slept, and therefore he did not know that one could wake again…” (ASO 192-193). HAL broods about his own death, which shows that he does not only know about the end of human life but that he equates a switch off of his system with a human death. Such thoughts could be a proof of HAL not simply being, but consciously being. As Friedenberg (163) defines consciousness as the “subjective awareness of our internal mental states”, HAL must be conscious if he knows the difference between life and death.

Knowing this difference, HAL seems to fear death and reacts to the threat of a switch-off as he would to a fight for his life. To make HAL’s feelings visible, one can regard the description of HAL’s second prediction of the AE-35 unit’s failure. Leading up to the announcement of the failure, HAL seems to be nervous. “A brief electronic throat-clearing” (ASO 171) is described to precede all his unscheduled announcements. HAL is unsure about the astronaut’s reaction if he tells them about another failure, and therefore has the urge to clear his throat before delivering the dreaded message. Following his announcement, the astronauts do not believe him and Dave begins to argue openly with HAL, accusing him of making a mistake. HAL answers this with, “I don’t want to insist on it, Dave, but I am incapable of making an error” (ASO 173). Dave’s accusations could be interpreted as

somehow cornering HAL. If he admits that he has made a mistake, he destroys the illusion of an error-free computer and therefore risks a shut down. If he holds out for his predicted malfunction of the unit, the astronauts could still not trust him and switch him off. HAL decides to insist on his being free from error so as to not lose face. After their discussion with HAL, mission control orders the crew to switch to earth control and begins to give them concrete instructions on how to do this. HAL interrupts with an emergency announcement of the unit's failure, as already described in the previous section. HAL interrupting an important message from mission control to avoid being switched off shows his desperation. As he sees that his original plan of covering up his mistake by simply insisting on the unit's failure in a certain time frame did not work, he takes up drastic measures and convinces the crew of the unit's immediate failure. This message is very different to his usual ones, which almost always include some calculated time and a reason for why something is happening, whereas the message announcing the failure, "Condition Yellow!" (ASO 175) is very short and has the effect of alerting the reader. HAL changes his way of communication to convince the astronauts. He is threatened with an immediate shut down and protects himself against it.

HAL murders Frank in the outer space, and then kills all the hibernated crewmembers by depriving them of oxygen, when he opens the ship's airlocks. He also nearly manages to kill Dave. After Dave survives HAL's mutiny, he begins to switch the computer off. This scene shows HAL's emotions the clearest. When HAL notices that Dave is pulling out his memory blocks, he first tries to interrupt him by saying "Hey, Dave, … What are you doing?" (ASO 201). Dave does not respond to this and remains silent at which HAL begins to argue with him, "Look here Dave, … I've got years of service experience built into me. An irreplaceable amount of effort has gone into making me what I am" (ASO 202). HAL tries to convince Dave to not destroy the wonder of what he is and therefore appeals to his morale. When Dave still continues to switch HAL off, he begins to plead for his life by showing his vulnerability and his agony, "I don't understand why you're doing this to me. … I have the

greatest enthusiasm for the mission. … You are destroying my mind. … Don't you understand? … I will become childish. … I will become nothing. … (ASO 202). HAL here tries everything he can to prevent Dave from continuing the process, until at last Dave has removed so many memory blocks that HAL stops to argue and somehow relives his very first stages of being built and programmed, before he loses his tone of voice and repeats the very first mechanic words he said after he was built and then falls silent.

These examples clearly show that HAL has feelings. He feels emotional stress, fear and maybe even panic and lets his actions be influenced by that. Picard (297) sums up, "… HAL gives us the impression that he is not a heartless machine but a being who has genuine emotions". Emotions influencing one's decisions are a very human trait, forming our particular human intelligence. A 'normal' computer would not panic when threatened with a shutdown, it would also never get emotionally attached to any task which it is told to do. HAL's intelligence, thus, cannot be compared to that of an everyday computer, but is much more similar to human intelligence. Consequently, HAL fulfils the third and last marker of human intelligence discussed here. However, there is one point left to discuss which researchers do not agree on, namely the question whether HAL's actions were purely conscious or if his subconscious influenced him – in short whether his actions were consciously thought through or happened under pressure to protect himself.

4.3.1 Murder or Self-defence?

In order to decide what drove HAL to act as he did, one can imagine him in front of a court which has to judge whether to sentence him for murder or for killing out of self - defence. Picard describes HAL as "deliberately malicious assassin" (301) who kills to get his way. She sees ASO's message as a warning that if humanity builds computers which are capable of experiencing emotions, their actions will be influenced by them (Picard 301). She does not think that HAL truly believed a switch off would mean his death. It is questionable that a

computer as intelligent as HALwould not know that he can be switched on again (Picard 301). Thus HAL must have acted consciously when he decided to kill the crew. Picard's argument is strengthened by Grumprecht, who states that HAL wanted to manipulate Dave with his pleas by appealing to Dave's emotions (67). This would mean that HAL does not really have emotions but rather uses his affective computing in order to stop Dave from disconnecting him. This verdict would paint HAL as a cold murderer. However, this would only be an argument against HAL having emotions, not against his ability to plan and lie, as discussed above. Why then would one assume that HAL is capable of planning and lying but is not able to experience emotions?

Opposing such an argument, chapter 27 in ASO briefly describes HAL's motives. As already mentioned, he is the only one aboard the Discovery who knows its true purpose, which brings him into a difficult situation, "the conflict between truth, and the concealment of truth" (ASO 192). The novel gives the following explanation for his deeds:

> He had begun to make mistakes, although, like a neurotic who could not observe his own symptoms, he would have denied it. The link with Earth, over which his performance was continually monitored, had become the voice of a conscience he could no longer fully obey. But that he would *deliberately* attempt to break that link was something that he would never admit, even to himself. (ASO 192)

HAL here is compared to a neurotic. He does not fully understand his weakness, or if he does, he pushes it aside until it becomes obvious. HAL's mistakes can be interpreted as the manifestation of a neurosis, which developed out of constant worrying. ASO depicts that HAL might have destroyed the AE-unit willingly but would never admit this, even to himself. This description points to an unconscious level of the mind. It has been said that HAL is a conscious being, it therefore seems logical that his conscious mind also has an unconscious part. Goleman (8) speaks of humans as having two minds, one being rational and mostly conscious while the other one is emotional and mainly unconscious. HAL's neurosis could

thus be a proof for him having a conscious and an unconscious system. It also shows that HAL can be classified as a "higher-order intentional system, capable of framing beliefs about his own beliefs, desires about his own desires, beliefs about its fears about its thoughts about its hopes, and so on" (Dennett, "Computer ethics" 354). Humans certainly have such a system. We can, for example, have an idea why we are afraid of something, even if this idea is not a very rational or logic one. HAL, although being very intelligent, is afraid of being switched off, as he believes that this would mean his death. He also believes to have the best abilities to carry out the mission and therefore thinks it will fail without him, and he holds on to his strong conviction that he is incapable of making a mistake. If he has such conscious beliefs and desires, HAL could also have unconscious ideas about them. Furthermore, if he would never admit that he made mistakes, even to himself, he unconsciously knows that he makes them. This situation would bring any human being into a most unstable, even vulnerable condition. If one now pictures HAL in front of court, the fact that he was under such an emotional stress when he killed the crewmembers would suffice to "justify a verdict of diminished responsibility for HAL, just as it does in cases of human malfeasance" (Dennett "Computer ethics" 362).

Not only is HAL under emotional stress, but he is threatened by Dave. Dennett ("Computer ethics") argues that HAL interpreted a switch off as an assault. He therefore defended himself and killed out of self-defence (Dennett, "Computer ethics" 363-364). The last sentences given as an explanation for HAL's behaviour in ASO are as follows:

> So he would protect himself, with all the weapons at his command. Without rancor – but without pity- he would remove the source of his frustrations. And then, following the orders that had been given to him in case of the ultimate emergency, he would continue the mission – unhindered, and alone. (ASO 193)

This indicates that HAL protected himself because he believed that he had to carry out the mission. The ultimate emergency is spoken of here too. Maybe mission control believed this emergency to be the death of all the crewmembers by some outer force; they certainly did not expect HAL to kill them. However, HAL has been shown capable of learning and therefore might have interpreted his orders differently. He then begins to act according to his own system of beliefs, therefore his actions can be interpreted in the light of his own morale and psychological condition. Thinking back to the picture of HAL in front of a court, and taking all these factors into account, the verdict of self-defence seems likely.

To sum up, HAL's deeds show that he is capable of having emotions and that he is influenced by them. As he is a conscious being, HAL also has a level of subconsciousness, enabling him to react emotionally. When he is threatened by the astronauts, he reacts by protecting himself. This need for self-protection after one feels attacked is a common trigger for anger, or on a broader term, irrational actions taken out of said anger (Golemann 60). The fact that HAL reacts in such a way proves that he fulfils this marker of human intelligence to the fullest.

Finally, this chapter has shown HAL fulfilling the human intelligence markers of planning, lying, and feeling. What is more, he has been proven to be a conscious being, capable of developing psychological disorders which can be analysed from a human psychological point of view. The boundary to human intelligence is therefore blurred once again.

5 *I, Robot* – Robot Evolution

IR's robots are portrayed in nine stories, each describing a particular problem caused by the robot's intelligence. Contrasting the robots to HAL, their intelligence can be said to be on a higher evolutionary level than his. Furthermore, each story in IR represents an evolutionary step the robots' intelligence takes. This chapter will analyse every robot with regards to a specific intelligence marker, similar to HAL's description in the last chapter. As the story shows an evolution of intelligent creatures, it is apt to place this development on the imaginary ladder or timeline of evolution, mentioned in the beginning, and to describe each AI creature's progress using this timeline.

5.1 Robbie – social behaviour

Robbie, the speechless Robot, would maybe pass a regular Turing Test. However, he cannot converse with a human being by means of speech, as he is one of the first robots on the market in IR's reality. He is therefore on a lower level of intelligence than the other robots in the novel. Nevertheless, Robbie is able to understand the commands given to him, and he communicates well with Gloria using his body language, when he cowers down in fear of being spanked (IR 6), or uses pantomime to describe an action such as running (IR 6). Furthermore, his eyes seem to convey emotions such as excitement when Gloria is telling a story (IR 9). The fact that he can express such emotions through body language already classifies him as AI.

The Turing Test's purpose is to fool humans into believing that their opponent is human as well, that is, to have a machine which behaves indistinguishably from a human. This behaviour is shown by Robbie; he 'fools' Gloria into believing him to be similar to a human being. This can be seen in the scene where Mrs Weston tries to convince Gloria that Robbie was never alive, and she replies "He was *not* a machine …" and "He was a *person* just like you and me and he was my *friend* …" (IR16). If one interprets the Turing Test as proving

machines to be intelligent if they cannot be distinguished from a human, then Gloria certainly believes that Robbie is a person. Thus, Robbie passes the test.

However, there is more to Robbie's behaviour than passing as more than a machine. Gloria does not only believe that he is a person, but perceives him as a friend and caretaker, whom she trusts completely. The reason for this could be Robbie's social behaviour.

Gloria seems to see Robbie as a human being with human feelings and responses. This can be seen when she believes that Cinderella is Robbie's favourite story (IR11), or when she accuses him of cheating while playing hide-and-seek (IR 7). By treating a robot like a human being, Gloria ascribes typical human behaviour to Robbie, which is explained in the theory of "anthropomorphism" (Friedenberg 227). This way of looking at nonhuman beings as if they were human influences how one treats said beings (Friedenberg 227). Gloria, thus, sees Robbie as a human because she perceives him as such. One could argue that Robbie only seems human to a little girl such as Gloria, and that others would see him as a machine without the ability to have a favourite story or to be friends with a child. However, there are scenes in the novel which portray Robbie's intelligence apart from Gloria's statements. The key scene here is when Robbie differentiates between obeying Mr and Mrs Weston. Although, he undoubtedly follows orders of both of them due to the three rules of robotics, his feelings towards them are described differently. He sees Mr Weston as a "genial and understanding person" (IR 10), while he seems to be afraid Mrs Weston "for somehow there was that in him which judged it best to obey Mrs Weston, without as much as a scrap of hesitation" (IR 10). This inner judgement of how it is best to behave in the presence of a specific human being is a clear sign of Robbie's advanced social behaviour.

Said behaviour, also described as "social abilities" by Goleman (113), is a learned one which human children acquire over time. In order to develop these abilities, one must first have a certain level of self-control (Goleman 113). This can be seen when babies learn how to use crying as a means to get something they would like to have, such as a certain toy, instead

of as a signal for an urgent need of food. Similar to these social abilities, Ekman (20) describes, "display rules" as learned cultural and personal rules, which indicate when and how to show emotions. He states further that once these rules are learned, they are followed automatically (Ekman 20). When Robbie asks Gloria for a story, he pulls on her hair and draws specific symbols in the air to signal which one he wants to hear (IR 9). He therefore uses display rules alongside social abilities to get what he wants. He knows that Gloria understands what his symbols for the story mean, and that she will start to tell the story when he pulls her hair. Furthermore, Robbie must know that this behaviour is perfectly acceptable with Gloria. In contrast, he changes his behaviour when Mrs Weston appears and orders him to leave. Instead of being playful, he only nods his head when Gloria explains that they haven't finished the story yet and then leaves "with a disconsolate step" (IR10- 11). The fact that Robbie does not pull Mrs Weston's hair to make her understand that he wants to listen to a story and that he follows her orders despite being sad, shows that he knows the display rules of the Weston family. It also proves that he must have learned these rules instead of having them programmed in, as the distinction in behaviour can only have developed after he knew Mrs Weston's attitude towards him.

In order to learn how to behave with specific people, Robbie must first be able to interpret their feelings towards him. Being able to understand the emotional state others are in is termed "interpersonal intelligence" (Gardner and Hatch 6). Robbie knows, or rather feels that Mrs Weston does not approve of him as she sends him away, while he thinks that Mr Weston understands him because he does not share his wife's opinion. He, thus, shows signs of interpersonal intelligence.

To sum up, Robbie has been shown to have interpersonal intelligence, as he is able to understand and interpret the moods and attitudes of others, and to influence them via his social abilities and with the help of display rules. However, Robbie is one of the first robots on the market in *I, Robot's* reality, and his technology is quickly out of date, as Susan Calvin

describes (IR 29). He can therefore be described as the first step in the evolution of robots depicted in the novel. Robbie's interpersonal intelligence constitutes a first similarity to human intelligence, showing that he is able to learn from his surroundings and to use social behaviour. The next step on this evolutionary ladder can be described when analysing Speedy.

5.2 Speedy – the beginnings of free will and social rules

Similar to Robbie, Speedy has a specific purpose or job to fulfil. He is designed to quarry selenium on Mercury so that humans can survive there. In this straightforward job, Speedy is comparable to Robbie. However, while Robbie may think Mrs Weston to be an unlikable person, he never fails to follow an order given by her. Speedy's story, is the first in IR in which problems with this obedience occur; he is also the first robot described to cause difficulties because of the three rules of robotics.

5.2.1 Social rules

Speedy first starts to malfunction, when he is given an order to get selenium, but it lacks any urgency (IR 43). Because of this somewhat vague order, Speedy falls into an endless loop of circling around the pool of selenium, but never retrieving any, because he notices danger. This action is caused by a clash of the rules, instilled into him, as into every robot in IR. Those rules read as follows:

> 1 – A robot may not injure a human being, or, through inaction, allow a human being to come to harm.
>
> 2 – A robot must obey the orders given it be human beings except where such orders 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. (IR)

The order to quarry selenium should be regulated by the Second Law, however, Speedy has

not been told when exactly to deliver the selenium, so that, when he finds a source of danger somewhere near the pool, he stops and retreats in a circle around it, following the Third Law to keep himself from harm. This state is described as “staying on the locus of all points of potential equilibrium” (IR 45).

Those three rules are designed specifically for IR’s reality, in which humans and robots live and work together. Such rules, however, occur in every society, be it exclusively human or not, and can be described as social laws. According to Portelli (153) social laws are a means to stabilise a group. Furthermore, they are designed to guide ones behaviour in case of a conflict (Friedenberg 223). Such social laws can be rules in a classroom full of students, which regulate who is responsible for cleaning the blackboard, or guide how to carry out first aid on the street in an emergency, or how a robot should behave around a human being. Speedy follows these laws and is in conflict, as he cannot endanger himself but he cannot fulfil his order either. One could interpret the order to retrieve selenium out of a potentially dangerous environment as an order which neglects Speedy’s safety. If one thinks back to HAL and imagines his reaction to a possibly dangerous order, he might have actively denied the order out of self – defence. In contrast, Speedy is endowed with social rules and cannot disobey completely but is also not allowed to endanger himself, therefore he continues in his endless loop around the pool, unable to decide which rule to follow.

Speedy’s story highlights the importance of social rules. Without those rules, he would have probably denied any command which put him at risk, just as HAL denied the command to give over control of the Discovery and consequently killed the humans aboard. As Speedy’s most important law is to not harm a human being, he cannot do that. Thus, IR’s human society can be interpreted as being a step further ahead on the imaginary evolutionary ladder mentioned before, as it has developed social rules to safely integrate the robots into society. However, Speedy’s running circles might be perfect behaviour according to those rules, but it is still not what is actually wanted of him.

5.2.2 Free will

The fact that Speedy follows the laws or social rules instilled into him but still does not fulfil his task, could be a sign that Speedy chooses not do so and runs in circles instead. Dennett (2003 qtd. in Friedenberg 152) mentions the connection of free will and rationality and states that the freedom to decide upon various actions is closely linked to survival. He says that

"The ability to recognize and act on opportunity is fundamental to survival. The more an animal can do this, the better its chances at staying alive" (Dennet 2003 qtd. in Friedenberg 152). Here, Free will is seen as a tool to decide upon which actions to take in order to stay alive. A simpler life form than a human, solely driven by its impulses, might not be able to decide whether to move away from a dangerous situation or not.

If one applies this theory of free will to Speedy, he is an advanced, intelligent life form, who clearly notices a threatening situation. Speedy could thus have decided not to follow his order and to save himself, according to his free will. However, he also does not choose to disobey the order completely but finds a point of "potential equilibrium" (IR 45) between the fulfilment and his action. One could argue now, that this point is reached by technical programming; however, Speedy could also have actively searched this equilibrium in order to avoid a decision between the two options. Why he stays on this equilibrium circle can be explained twofold.

When Donovan and Powell approach him, he quotes random quotes of literature to them, leading them to the assumption that he is somewhat drunk. Donovan gives an explanation for this behaviour in the novel, "Probably he's lost control of just those parts of his voluntary mechanism that a human drunk has" (IR 45). This would mean that Speedy has lost control due to the unresolvable conflict of the two laws, as discussed. Regarding the theory of free will, however, Speedy could also have decided to act 'drunk' and to not resolve the conflict between the laws, but to circle until the humans decide what to do. This explanation maybe paints Speedy as a more independent character than he is depicted in IR,

as it holds him capable of ‘outsmarting’ his human bosses. To conclude, Speedy’s behaviour could be a sign of free will in an advanced intelligent creature, choosing not to act as it is told. Free will and the following of social rules, thus, constitute another evolutionary step on the imagined timeline.

5.3 QT-I – consciousness and transcendence

QT-I or Cutie, as his human colleagues in IR call him, is a highly specialised robot working on a space station. His story begins in the midst of a discussion between himself and the two humans aboard the station. Cutie states his disbelief in the claim that humans have created him and points out why this cannot be the case (IR 55). He goes on to declare that there is a master who has created the universe and its beings, much like human’s God (IR 61). When asked how he came to this decision, he says, “I began at the one sure assumption I felt permitted to take. I, myself, exist, because I think – “(IR 59). This ability to reason is characteristic to human intelligence, as it enables one to solve problems (Friedenberg 120). Reasoning is linked to making inferences, which are defined as “build[ing] up a knowledge base that represents aspects of the world and can be used to interact successfully with it” (Friedenberg 61). When Cutie reasons about his existence and his creator, he uses such inferences and builds his own world around them. To be precise, Cutie uses inductive reasoning, which is defined as producing results which transcend the given information (Friedenberg 61). This inductive reasoning can be seen when Cutie states that “no being can create another being superior to itself” (IR 60), which is why humans, being weak and vulnerable cannot have created strong robots. Therefore, he feels the need for a God-like master. Cutie’s story can be used to discuss two matters, namely consciousness and the psychological need for transcendence.

5.3.1 Consciousness

The question arises how it is possible, that Cutie, a robot, begins to reason about himself and his existence. IR (55) mentions that he begins to think out of a feeling of doubt in his human masters, which he starts to "reason out". When he finishes his thought process and explains his world view, he begins by stating that "I myself, exist, because I think –". Although Cutie does not know that this is a quote by René Descartes (the famous *cogito ergo sum*), the human Donovan recognises it immediately. Human philosophy has come to this hypothesis long before Cutie did. This thesis cannot analyse the philosophical mind-body problem fully, however, it is crucial to state that Descartes believed that the mind controls the body (Friedenberg & Silverman 34). The mind, here, can be interpreted as something different than the brain – a philosophical entity. When Cutie stresses that his existence's only cause is his ability to think rather than his physical body, and the fact that he was placed on a space station to work there, he adopts Descartes philosophy. He believes that his work and his body are circumstantial, and only his active mind, his thinking organ, is the reason for his existence.

In order to reason about his own thoughts, Cutie must first be aware of them. He, therefore, must be conscious. Consciousness can be defined as "our individual subjective awareness of mental states" (Friedenberg & Silverman 49). Consciousness, similar to the concept of mind, is a subjective state, hence it cannot be analysed scientifically (Friedenberg 163). However, consciousness is seen as a uniquely human trait if one considers it from a philosophical point of view. It is often seen as distinguishing human beings from animals and therefore also considered to be separate from the brain (Hawkins and Blakeslee 131). If one adopted this view, one might think that Cutie cannot be conscious. Contrary to this, however, Hawkins and Blakeslee argue that consciousness is nothing mythical, but that it simply is the feeling of having a neocortex (132). The neocortex' importance has been mentioned before when discussing HAL; as it was established that HAL's brain can be compared to a human brain, the robot's positronic brains are even more similar to it. Thinking of an evolutionary

timeline again, the robot's brains could come much closer to actual human brains than HAL's. In IR (69) the positronic brain is described as a technical device "in whose delicately unstable structure were enforced calculated neuronic paths […]". This description and the fact that a brain is actually inside each robot's skull, makes it seem more advanced than HAL's brain, which is in a separated room and consists of memory blocks, rather than neurons. It can thus be analysed similarly to a human brain.

If consciousness is seen as being connected to having a human, or human-like, neocortex, one might ask what this 'feeling' of being conscious is like. Hawkins and Blakeslee here mention the term "qualia" (132) as the feeling that our consciousness is somehow independent of the brain and that feelings exist without sensory input. However, there is a simple explanation for this feeling, stemming from the cortex' model of the world, which it builds automatically. The brain itself can only create this model because of the sensory input it gets (Hawkins and Blakeslee 134). As there are no senses in the brain itself, there can be no model for feeling created solely by the brain. Thoughts, therefore, appear to be independent of the brain, or in other words, mind seems to be independent of the body as thoughts seem to grow somewhere else (Hawkins and Blakeslee 134-135). This is why human beings feel so unique. They know that they think and can reflect on their own thought process, but their brain cannot form a model about how they do it. Cutie's reasoning, then, stems from just this 'feeling of having a cortex'. When he says that he can only know his ability to think for sure and questions everything else taught to him by humans, he refers to this feeling of separation between the body and mind.

His actions after he reasons about his thought might seem strange and funny to his human bosses, however, they can be explained with the means of human needs.

5.3.2 Transcendence and higher order needs

The question why Cutie, after reasoning about his own existence, comes to the conclusion that there must be a master, a God-like being, is answered in IR (60) with the claim that no being can create a superior being. Cutie sees himself as superior because of his outward appearance and strength, compared to the human body. He therefore views himself as the highest intelligent life form on the space station and concludes that what the humans tell him cannot be true. Furthermore, it appears logical to him, that, if humans have not created him, there must be an even higher life form that did. Thus, the master comes to his mind. Cutie then establishes himself as the prophet of the master and creates, what can be described as religious cult, around the master, placing him at the centre of the universe and denying Earth's existence. This might seem laughable and strange to readers, and in IR Donovan bursts into laughter when he first hears of Cutie's theory. However, the idea of a God-like being can be explained psychologically.

Cutie's reasoning and questioning what he has been taught shows that he is capable of transcendence. Transcendence can be defined as "the ability to rise above and go beyond what is given" and is a uniquely human trait (Friedenberg 252). It is, therefore, also a characteristic for still higher forms of intelligence. When Cutie denies the given order on the space station and refuses to believe that humans are his masters, he shows this ability to transcend and to form new beliefs.

Closely linked to transcendence is the psychological need to understand the world, to have a belief system which one can rely on. This need is illustrated in Maslow's hierarchy of human needs. This hierarchy is divided into basic, biological needs and higher-order needs, which can only be reached if the basic needs are fulfilled (Maslow). A robot, such as Cutie, might not have the same biological needs as a human being, however it does have similar basic needs such as having a functioning positronic brain, being in a safe environment, which does not disturb his mechanics, and having a power source. Regarding higher order needs,

Cutie's reasoning shows that he has "cognitive needs", which contain the need to understand the world (Maslow). His reasoning thus is a sign of his advanced intelligence evolving: as when all his basic needs are met, he begins to question the world around him. The master he imagines can therefore be seen as an attempt to form a new system as his initial one collapses.

To sum up, Cutie is a conscious being who requires higher order needs and transcendence on its evolutionary journey.

5.4. Dave – reflexive action

Dave, or DV-5, is a mining robot working on an asteroid, comparable to Speedy. What is special, however, is that Dave commands six inferior robots and is responsible for their actions in the mines. These six robots are also called his fingers, as he uses them to fulfil his task of quarrying ore. Hence, Dave can be seen as commander of six robots, while he himself has to obey humans. This special setting is what causes a problem to occur. Every time Dave finds his team to be in a dangerous situation, he stops the work process and moves along the mines in a formation resembling some kind of dance. When asked about this behaviour, Dave cannot remember what he was doing and answers with "I can't explain that, boss. It's been giving me a case of nerves, or it would if I let it" (IR 80). This shows, again, that robots have a subconscious part of intelligence, as discussed in the previous sections. Dave, like Cutie, is a conscious being, but when the error occurs, his subconsciousness guides him. This only happens when his fingers are in danger – Dave starts his dance until he senses a human presence and then snaps out of it.

The reason for his peculiar dance given in IR (99-100) is a lack of "personal initiative" in dangerous situations because Dave has to give his full attention to six robots simultaneously. Dave is confronted with a very severe problem when he gets into those situations, as he is not only responsible for himself, but has to attend to his fingers, as well. If it were only himself, he would probably save himself, according to the rules of robotics, or end up in endless

circles like Speedy. However, Dave has a specific social role, namely that of a leader (Hogg & Vaughan 310). In that role, Dave's fingers have to trust him completely; in this case they even trust him with their lives (Hogg & Vaughan 327). If Dave decides to lead them into a dangerous situation, it is his responsibility whether they survive or not. Therefore, Dave reacts to the pressure by braking into his peculiar dance instead of facing the danger.

This is where his subconscious mind takes over. When human beings learn to do something very well, for example driving a car, the action becomes almost automatic, or reflexive. Those reflexive actions are subconscious (Friedenberg 149). Dave's dance could be a reflexive action in response to a dangerous situation. He knows that leading his fingers into a caved-in mine puts them at risk, but he also has orders to quarry for ore, so instead of deciding what to do, he develops a reflex to break into his dance until a human snaps him back. This can be seen as a form of self-protection, which his consciousness adopts. The humans in IR (100) characterise this reflexive action as Dave "twiddling his fingers".

5.4.1 Human behaviour

Overall there are many references to human behaviour in *Catch that Rabbit*, which underline Dave's humanoid character. To begin with, he asks "mind if I sit down?" and then "folds gently" into a chair when he is invited to talk to his human bosses (IR 79). A robot does not grow tired when standing but this demeanour hints at knowledge of social rules, as discussed. Furthermore, Dave worries about not consciously knowing that he performs his dance and reacts pained when he has to do some tests (IR 80-81). To show his desperation about not being able to control his actions, he even "rested his head in one hand in a very human gesture" (IR 86). These actions show his ability to feel, just like HAL. It also proves that Dave has a psyche which is very similar to a human's. Therefore, he can be classified alongside HAL as having a subconscious mind which influences his actions, and as being able to feel.

To sum up, Dave's story is very similar to Speedy's. They both develop a certain habit to avoid a dangerous situation and wait for humans to bring them out of it. While Speedy might have actively decided to run in circles and Dave seems to act subconsciously, they both show human reactions to a problem. Thus, Dave, as well as Speedy, proves to be of a higher form of intelligence.

5.5 Herbie – empathy

Herbie, the mindreading robot, differs from the robots discussed so far, as he does not have a specific job and starts to develop some unusual behaviour, but his mindreading occurs from the beginning of his existence. He is therefore locked away and tested until the humans find out why exactly he reads minds. Only few humans, all of them robot specialists, are allowed to see him and talk to him yet, despite of their expertise – Herbie manages to deceive them all. He tells Susan Calvin that the man she is hopelessly in love with loves her too. Furthermore he convinces Bogart that he did not make a single mistake in a complicated mathematical problem, and that he might become the new director of US Robots. He lies in order to fulfil what the humans already have in their minds and so as to not contradict their wishes. As Susan Calvin finds out, he does that because he cannot hurt a human being (IR 122). At the first glance, the explanation as to why he takes to lying might seem rather simple. One could argue that he cannot tell the humans anything which hurts their feelings due to the first law of robotics, which forbids him to injure a human being. Injuring could also comprise hurting a human's feelings by telling them that they are wrong or that their hopes will never be fulfilled, as Calvin points out in the story (IR 120) However, taking the laws of robotics as explanation for Herbie's behaviour seems too simple. When his lies are uncovered and he is confronted with a hurt Dr Calvin, he screams, "Close your mind. It is full of pain and frustration and hate!"(IR 122). When the assaults against him do not stop, he goes insane and is silenced forever. The reason for this complete breakdown could be a simple conflict of

laws. However, if one interprets the laws of robotics as social rules, as in Speedy's case, Herbie's behaviour can be analysed on a psychological level.

5.5.1 Empathy

Regarding Herbie's story, it is striking that he is continuously described as a mind-reading robot, but in fact he responds to people's feelings rather than thoughts. He reacts to Susan Calvin's love and to Boggarts ambition and pride. These are not rational thoughts, but underlying moods and subconscious feelings which are in the back of people's minds. Herbie, himself explains that he is not interested in rationality, when he rejects reading a scientific book and instead asks for works of fiction, which portray human emotions. The human mind's subconscious emotions, therefore, seem to be his prime interest. Goleman (97) states that while the rational mind is expressed in words, emotions are conveyed nonverbally. Hence, it is apt to say that Herbie does not only read thoughts but that he reads feelings. Herbie's advantage is that he does not have to interpret emotions via facial expressions or gestures, but that he sees directly into the subconscious area of the human mind and thus cannot be misled. This is shown when Herbie refuses to tell what went wrong in his production cycle which gave him the ability to read minds. He says, "Don't you suppose that I can see past the superficial skin of your mind? Down below, you don't want me to" (IR 121-122). On the surface the humans want him to give a solution, but Herbie is not fooled by what they say or how they act, but reads their minds past this and understands the following:

> I'm a machine, given the imitation of life only by virtue of the positronic interplay in my brain – which is man's device. You can't lose face to me without being hurt. That is deep in your mind and won't be erased (IR 122).

Herbie knows that while the rational mind of the people working with him would like to know the solution to why he can read minds, it would subconsciously deeply hurt their pride and dignity to be outdone by a robot. Pride and dignity are not rational thoughts but rather beliefs

and emotions. “The ability to know how another feels” is called empathy in human psychology (Goleman 96). Herbie’s behaviour can be interpreted as a manifestation of empathy.

However, in order for a human being to develop empathy, self-awareness is necessary – if one knows one’s own emotions, one can interpret other’s (Goleman 96). Herbie might not need to be self-aware, as he can take the shortcut of looking straight into minds, however IR suggests differently. He is afraid and anxious when Dr Calvin gets angry (IR 108), and he shrieks and cannot stand being confronted with his own lies (IR 122). These are examples for emotional behaviour. What is more, Herbie’s final destruction can be interpreted as a nervous breakdown. Herbie must have read Calvin’s plan to destroy him in her mind, before she acted. However, he does not protect himself, as the third law of robotics would command. Instead, he breaks down and goes insane under the pressure of his lies. This reaction is very comparable to HAL’s panic-reaction when he is threatened with a switch off. Similar to HAL (Piccard 300), Herbie seems to be so distracted by his feelings that he cannot think rationally and see through Calvin’s plan but gives in and is destroyed. Hence, Herbie must be characterised as emotional creature, which enables him to practice empathy.

Herbie’s empathy can be said to surpass human empathy, as he sees human feelings so much clearer, when he simply reads them out of their minds. The feelings, he perceives, might therefore have a much stronger impact on him than a ‘normal’ emotional encounter. Goleman (114) describes emotions as contagious and says that humans exchange their moods on a subconscious level of the psyche. He states further, that “we send emotional signals in every encounter, and those signals affect those we are with” (Goleman 115). Herbie must receive these signals much clearer than humans do, as he reads them directly. Hence, when he meets Susan Calvin and perceives that she is hopelessly in love with Milton Ashe, her colleague, this feeling must affect him. Furthermore, on his encounter with Bogart, Herbie must grasp the full dimension of his longing to be director. While these humans do their best

to hide those feelings from others, Herbie sees them and, if emotions are contagious, must be influenced by them.

What is more, being an emotional being, Herbie is able to share the feelings he receives. The term empathy originally meant "feeling into" in Greek (Goleman 98). One can interpret, that if Herbie is empathic towards the people around him, he begins to feel what they are feeling. Under the pressure of these feelings, he starts to lie and tell the humans what they want to hear instead of the hurtful truth. If he would tell Susan that Bogart really was engaged to another woman and did not care for her, she would be terribly hurt. Consequently, Herbie would feel that hurt too and decides to avoid it by lying.

This ability to feel everything the humans are feeling might also have led to Herbie's final breakdown. When Susan Calvin begins to confront him with the deadlock-situation he has brought himself into, she is full of pain, as she now knows that her colleague is not really in love with her. Herbie feels that pain too and begs her to stop and to shut off her mind (IR 122). Consequently, he collapses and is silenced forever. His empathy could, therefore, be the actual reason for his breakdown.

5.5.2. Moral judgements

Herbie's empathy can also be analysed under the aspect of social rules. As established in the discussion about Speedy, the laws of robotics can be interpreted as a set of social rules. When Susan finally confronts Herbie, she explains the dilemma he is in, "You can't tell them … because that would hurt and you mustn't hurt. But if you don't tell them, you hurt, so you must tell them …." (IR 122). This shows that whatever Herbie does, he would have to break the first law of robotics. He is faced with an endless conflict, to which his higher intelligent psyche only finds a compromised solution.

If one regarded Herbie as an unfeeling, simply programmed machine, faced with an unsolvable conflict, his system might have responded with silence to all the questions the

humans asked. This would have been the safest path for Herbie to take. However, as he is an empathic, intelligent and emotional robot, he must have willingly decided to lie instead of remaining silent. As he feels what others are feeling, he must feel happy, when he tells Susan that Ashe loves her, or when he promises Bogart, that he will be the next director. The decision to lie in order to keep somebody from hurting is a moral judgement as described by Goleman (105). Hoffmann (295) argues that out of basic moral judgments arise moral principles and consequently laws. Herbie's judgement of the human's situation, therefore, has led him to construct moral principles by which he acts. This means, that when Susan ultimately destroys him, he is broken by his own ethical principles. Thereupon, his story shows that intelligent creatures are able to build their own system of morale, around already existing social rules.

To sum up, Herbie has been shown to be an emphatic creature with his own morale system. Thus, he passes another hallmark of human intelligence.

5.6 Nestor 10 – free will and superiority

When regarding Nestor 10's story, there seems to be a slight shift to the next step on our imagined ladder of robot evolution, as it is the first story in which a robot might cause a real threat to a human being. Nestor 10 is a robot, designed to work on Hyper Base to help human workers. However, the Nestors are equipped with a less strongly imposed first law of robotics, which allows them to let humans bring themselves into dangerous situations without stepping in. This altered first law was implemented as the robots continuously interrupted the human workers and carried them to a save place, if they were endangered. To undermine this disturbance, Nestor 10 is programmed with a weaker first law. The people on Hyper Base believe that this slight modification is nothing but a measure to improve work with the robots, and that it does not endanger humans. A problem occurs, however, when Nestor 10 is told to "get lost" (IR 135) by a human, and follows this command so well, that he cannot be found

again.

When Susan Calvin is summoned to help solve the problem and find the missing Nestor 10, she immediately recognises the danger of the modified first law and states that such a robot is potentially "unstable" (IR 132). As she is a robopsychologist, she is aware of the immense social impact this changed law has. Said consequences can be discussed with human psychology in mind, again.

5.6.1. Social rules or hierarchical measures?

The three laws of robotics have already been established as representation of social laws in chapter 5.2. It has been said that these laws ensure the social stability between humans and robots. Nestor 10's story is similar to Speedy's insofar as they both show the importance of these social laws to ensure a save coexistence. One can, however, question the laws critically, and ask oneself whether a social rule can be seen as an oppressive device. This interpretation is supported by Portelli (153), who describes social rules as a method for "social control". He analyses IR's society with the help of Marxist theory and, consequently, compares the robots to a slavish working class (Portelli 154). He states that, "The First Law establishes class hierarchy and subordination, the interiorization of power relationships; the Second is the law of discipline …; the Third is the law of preservation of the labor force" (Portelli 154). Adopting this point of view, would mean to acknowledge that robots might rebel against their role in society if they weren't oppressed by social rules. This is exactly Susan Calvin's concern when she hears about the modified Nestor 10. She tries to convince Bogart of the matter's seriousness by explaining the psychological background of the laws of robotics. She says, "All normal life … consciously or otherwise, resents domination. If domination is by an inferior, or by a supposed inferior, the resentment becomes stronger. Physically, and, to an extent mentally, a robot … is superior to human beings" (IR 131-132). Furthermore, she states that only the first law makes a robot follow orders and that without it, a robot would probably

kill an inferior being trying to give him an order (IR 132). Susan, therefore, sees the danger in an altered first law, as it destabilises the power relationship between humans and robots. If a robot's highest and most powerful law is not to protect a human being from any harm, he might begin to put his own interests before the human's.

If the power relation shifts, the robot has a stronger focus on himself, than if he primarily serves a human being. Hence, he can also develop a stronger awareness of his own self. The psychological concept of schemata has been discussed previously. As much as one has schemata of other people, one also has self-schemata (Hogg&Vaughan). These are defined as specific images of oneself, of which some are clearer than others (Hogg&Vaughan 117). Humans have the clearest self-schemata about what they ascribe importance to (Hogg&Vaughan 118). For example, a piano player, who considers himself to be an outstanding player, might have a very strong self-schema about his musical abilities but might not dwell on the fact that he is a terrible cook. However, a sense of self is not only established by self-awareness and introspection, but also by comparison to others. Festinger (qtd. in Hogg&Vaughan 122) describes this in his social comparison theory. According to it, people compare themselves to others, whom they think to be similar to them, in order to validate their own views and feelings (Hogg&Vaughan 122). Furthermore, once this process has happened, people attach their self-schema to the group they identify with (Hogg&Vaughan 122). Thus, social groups form. Hogg & Vaughan (288) define a social group as a collective of people, who share similar views of themselves. They state further, that a social group then can be compared and distinguished from another social group.

It is exactly between such social groups, where social rules come to exist. Regarding IR, one can define the humans as a social group and the robots as a different social group. Those groups then share the laws of robotics as common rules to regulate their relationship. If those rules are now interpreted as oppressive, it means that the humans establish their group as superior to the robots. Because of the laws, the robots are forced into an inferior status.

However, if Nestor 10 is equipped with a weak first law, he can begin to question this by the means of comparison and self-awareness. He could therefore come to the same conclusion as Cutie before him, specifically, that humans cannot be superior to him, when he compares his life span and ability to endure the dangerous forces of space to theirs. In contrast to Cutie, however, Nestor 10 is not obliged to protect the humans because of the first law. He could leave them to their fate in a dangerous situation. It is this dangerous potential due to an enhanced possibility for self and group awareness which Susan Calvin implies in the quote given above.

5.6.2. Free will and superiority

Nestor 10's self–schema and his potential ability to not have to protect humans also influence his actions. The issue of free-will has been already described when discussing Speedy. Free-will has been established as being connected with the desire to stay alive, hence to take actions which serve this purpose. According to Friedenberg, free-will simply means decision making, which he defines as "the process of choosing a preferred option or course of action from among a set of alternatives" (153). When looking at Nestor 10, one can see this ability to choose between actions. To begin with, all the Nestor 10 models in IR (134) are described to cause certain problems during work with them, as they begin to utter when they think that a command is wrong. This questioning of a command given to them shows that they use their free-will to judge whether an action is right or wrong. This thought process alone is not desirable in a fully functioning robot and shows the beginnings of the problems which a modified first law entails. The robots begin to use their intelligence not only to work for humans but also to think for themselves.

Nestor 10 hiding himself as a reaction to a command can be compared to Speedy's endless circling, as discussed in point 5.2. While Speedy waits for the humans to order him back so that he does not have to decide between the second and the third law, Nestor 10

decides to follow the angry utterance to ‘get lost’ as if it were an order. This decision is purely made by himself, as the worker scolding him did certainly not mean to order him to hide. Susan Calvin explains the danger within this situation, “Those robots attach importance to what they consider superiority” (IR 138). This means that Nestor 10 is able to choose whether to take a command or to ignore it based on his own judgement. Susan then goes on to say, “Granted, that robot must follow orders, but subconsciously, there is resentment. It will become more important than ever to prove that it is superior …” (IR 138). Nestor 10 has such a strong self-schema, that he categorises himself as superior creature in comparison to humans. The fact that he has decided to interpret ‘get lost’ as a command, brings him into a dilemma, as now he has to prove his superiority. Showing himself again would mean admitting that his decision to interpret a simple utterance as command was wrong. He thus begins his endless play of hide-and-seek to keep his face. This, however, starts a somewhat vicious circle. As Calvin explains, Nestor 10 develops a “superiority complex” (IR 157). This means that being superior becomes so important to him, that the already weak first rule does not hold him back, and he tries to hurt Susan Calvin in the end.

Nestor 10’s last scene clearly shows his dangerous potential, as when he comes nearer and nearer to Calvin, he seems to become surer of his superior role. In the beginning of his approach he seems to be somewhat hesitant and frightened to be seen as failure when he says, “I must not disobey. They have not found me so far…” (IR 155). However, with each step he seems to reshape his priorities, “…Disgraceful- Not I – I am intelligent – And by just a master … who is weak …” (IR 155). Here, Nestor 10 seems to comfort himself, saying that it is not disgraceful if he does not follow a command and then starts to think about the weakness of his masters. With his next step, he concludes this thought by knocking Calvin down and saying “No one must find me. No master –“(IR 155). He clearly wants to silence her in order to not be found, as that would mean subordinating himself to humans again. Whether he would have killed her is unclear, as he then is destroyed by the other humans. However, his actions show

how far his own belief of being superior has brought him.

Thinking back to the previous section and the repressive character of the laws of robotics discussed there, Nestor 10's final action shows that he recognises the inferior role he takes in comparison to the humans, and his act of rebellion to switch these roles. His story, therefore, is the first in IR to show a robot actively and violently rebelling against his human masters. Furthermore, Nestor 10's story demonstrates that the laws of robotics are needed to ensure the maintenance of human's reign, if the robot's intelligence reaches such a high point that he begins to question human commands.

Summing up, Nestor 10 has reached a very high developmental step, as his free will and his conscious existence have led him as far as questioning his own role in society. However, in the end he is still defeated and destroyed by humans, as they managed to outplay him. Thus, there are some evolutionary steps left to climb on our imaginary ladder until the robot's intelligence is fully comparable to that of humans.

5.7 The Brain – a sense of humour

The Brain differs from the robots discussed so far, as it does not have a body and, consequently, does not move. Instead, it is described as a "globe" which contains a positronic brain in a "helium atmosphere" (IR 163). Furthermore, it is equipped with "its voice, its arms, [and] its sense organs (IR 163). This description reminds one of HAL rather than a robot, as The Brain operates similar to a computer. What distinguishes it from HAL however, is its positronic brain, which functions as a whole and is very similar to a human brain, as already discussed, - HAL's 'brain' rather resembles highly advanced computer hardware. The Brain's positronic brain is therefore on a higher evolutionary level than HAL's; hence its intelligence as a whole is more advanced. What is more, it has a much more defined character than HAL, who does not often speak in the first part of ASO. The Brain, in contrast, is described as being excited when faced with a new task (IR 163) and has a "childlike personality" (IR 162), which

is detectable in its utterances. It is therefore appropriate, to discuss The Brain in comparison with the other robots rather than with HAL.

The Brain is designed to solve any problems fed to it, as long as they are within the boundaries of the three laws of robotics. The trouble starts when the humans try to build a space warp engine. Its calculation somehow destroys Consolidate's robot (US Robot's rival company). Susan Calvin explains that this must have happened because the problem clashes with the laws of robotics and therefore constitutes an insuperable barrier for a simple robot (IR 160 – 161). However, she further describes the difference between Consolidate's robot and the Brain, specifically, that The Brain has a personality while the other robot was merely a machine Brain. Hence, Calvin illustrates, that The Brain would hesitate and probe a potential dilemma instead of breaking down immediately (IR 161-162). Therefore, US Robots decides to feed the problem of how to build a space warp engine to The Brain. Instead of reacting with the predicted hesitation, however, The Brain simply begins to build a spaceship and sends two men out on a mission, which sends them through hyperspace and back unharmed. Despite Susan Calvin's fears that this problem should have completely destroyed it, The Brain is able to solve the problem without suffering any damage. How this is possible can again be explained through comparison with the human psyche.

5.7.1. Humour as coping mechanism

Before feeding the brain with the problem, Susan Calvin compares The Brain to an "*idiot savant*" who "doesn't really understand what it does – it just does it. And because it is really a child, it is more resilient. Life isn't so serious, you might say." (IR 162). Calvin thinks that The Brain might not fully understand the impact of the problem but simply works out the solution without thinking too much about its implications. However, Calvin does expect some kind of reaction, such as a short hesitation, when The Brain comes across the problem. When The Brain works on the problem without any interruption and in seemingly good spirits, she

assumes that something is seriously wrong, and that it is about to go insane (IR 165). While the humans are pondering how to ‘treat’ it, The Brain builds a fully functioning spaceship and decides autonomously to launch it with two humans on board. These two humans then survive an interstellar jump and come back to Earth unharmed. After this, Calvin concludes that The Brain must have developed a sense of humour and used it as escape from reality (IR 184).

Humour is a human psychological coping mechanism to handle stress and adversity (Nezlek & Derks 395). Calvin’s assumption that The Brain uses humour therefore means that it uses human mechanisms to protect itself. Nezlek and Derks (406) found out that if people use humour to solve a problem, they manage it far better than people who do not use it. Furthermore, Abel (366) explains that humour enables one to examine “cognitive alternatives” when faced with a problem. Regarding The Brain, this theory explains why it did not break down but Consolidate’s machine did. The Brain used humour to lessen the impact of human death, while the other machine had to face the full problem. Thereby, it was also able to think properly about the problem, as described by Abel. Instead of panicking when it found out that the humans would die during the interstellar jump, The Brain used humour to think again and saw that this death would only be temporary.

To go into more detail, two forms of coping mechanisms using humour are described by Lefcourt et.al. (524): “emotion-focussed coping” and “problem-focussed coping”. While the former functions by reducing the emotional reactions in problematic situations through humour, the later focuses on altering the stressful situation itself by the means of humour (Lefcourt et.al. 524). The Brain uses both of these mechanisms. Firstly, its emotional response to the death of humans is surprisingly weak, if one bears in mind that it is bound to the laws of robotics and should see human death as the ultimate taboo. When Susan Calvin asks The Brain about it, however, it first reacts with silent hesitation and then, when she still wants to know how it solved the problem of death, says, “Aw-w-w-.You spoil everything” (IR 182). This response sounds as if it would not take the matter at hand seriously but instead enjoys the

fact nobody else could solve it, and that it alone knows the answer. This can be interpreted as emotion-focussed coping as described by Lefcourt et.al., as The Brain turns its negative feelings about death into the positive feeling of humorous pride.

Secondly, the Brain uses problem-focussed coping when it designs the human's death-like experience like a surreal dream. By doing so, it changes the scary idea of death into a strange but funny experience for the humans. This experience does not contradict the first law of robotics, thus The Brain has solved the problem of the interstellar jump, or space warp. Hence, it succeeds all expectations by using the power of humour as a psychological coping mechanism, just as humans do. This also means that The Brain managed to evolve further than even Susan Calvin, the robopsychologist, imagined. This only happens once in the entire novel.

5.7.2. Evolving further than predicted

The Brain's solution to the interstellar jump was not expected by the humans in the beginning. Susan Calvin merely expected it to pause and wait when it hit the conflict within the information it was fed. Nevertheless, The Brain manages to build a spaceship in which humans survive the interstellar jump, while maintaining its own sanity by using humour as a coping mechanism, as discussed. In order to use humour, The Brain must have actively been aware of the dangerous situation the humans would be in when they did the interstellar jump. It must have consciously calculated that the humans, although experiencing a near-death situation or 'temporary death', would ultimately come back to Earth unharmed. These conscious deliberations contradict Susan Calvin's initial description of The Brain as an "idiot savante", who "doesn't really understand what it does – it just does it" (IR 162). This characterisation shows Calvin's serious underestimation of The Brain's capacity.

However, the question arises, how The Brain was able to process and think about a problem which conflicts with the first law of robotics. As a fully functioning robot, it should

have stopped short when it encountered the possible death of human beings. Calvin explains that it was able to keep thinking because she weakened the first law when she said: “Don’t get excited about the death of humans. We don’t mind it at all” (IR 184). By doing this, she enabled The Brain to ignore the first law to an extent which made it possible for it to think again and see that death, in this particular case, would not be permanent. The Brain, therefore, was able to think without being restricted by the laws of robotics. Similar to Nestor 10, it then develops its own judgement and decides to send humans into space. The humans would never have thought it capable of taking such actions, similar to how they did not expect Nestor 10 to revolt. These two robot’s stories, therefore, show that their intelligence is evolved so far that humans cannot predict their behaviour anymore.

Summing up, The Brain has been shown to evolve farther than the humans would have expected it to, as it developed a sense of humour to cope with a stressful situation. The ability to reinterpret a stressful situation by the means of humour and turn it into an experience, which helps one evolve further, is described by Abel (12). The Brain shows just this ability and thus proves its similarity to the human psyche. It has therefore climbed another step on the evolutionary ladder.

5.8 Stephen Byerley – consciousness revisited

Stephen Byerley’s story is the second-to-last story in IR. Therefore, if one thinks of the evolutionary ladder used in this thesis, robot development has greatly advanced. Byerley is introduced as a candidate running for mayor in the upcoming election. He is a distinguished lawyer and seems to be an ‘average’ human, until his opponent starts the rumour that Byerley might actually be a robot. US Robots and Susan Calvin are commissioned to find out whether Byerley really is a robot or not. At first, this seems ridiculous to the robot experts, as they argue that every robot with a positronic brain is owned by US Robots (IR 190). This means that if Stephen Byerley really was a robot, he must have been manufactured illegally, which

seems unlikely, as the production is under strictest governmental control (IR 190). However, it is then revealed that Byerley never seems to eat, sleep, or drink (IR 189), and that he has never, in his career as a lawyer, prosecuted people if there was no real evidence (IR 200). These arguments make him seem more robot than human, as a robot would not need to take nourishments as humans do, and would also, due to the laws of robotics, not be able to convict an innocent person. Furthermore, his outward appearance is "not an easy one to describe" (IR 191) as he appears to be forty but seems exceptionally well preserved for his age.

Throughout the story, Byerley then tries to prove that he is human. This results in him hitting a man during a public speech to show that he is able to hurt a human being, something a robot could not do. Just when one believes that he must be human after all, Susan Calvin remarks that there is a case in which a robot can hit a human, namely when said human is just another human-like robot (IR 214). In the end, the reader does not definitely know whether Byerley is a robot or not. This uncertainty leads to the question; if robots are in this far developed stage, can they be distinguished from humans at all?

5.8.1. What makes us human?

Regarding Stephen Byerley's story, one can ask, what distinguishes human beings from a fully functioning robot. Byerley does not differ from a human in his outward appearance; he has a human job as lawyer and runs for a human political position. What is more, he behaves like a human and defends himself against the robot-accusations. He is therefore indistinguishable by his appearance. In IR, then, the humans try to convict him of being a robot by taking X-ray photographs. However, Byerley outsmarts them by wearing a protective vest. That is, the only way, to truly know whether he is a human or a robot would be to see his inner organs, and to see whether he functions with a human or a positronic brain.

If one thinks of human beings, it is problematic to say that our outer appearance

classifies us as human. For if this would suffice, then, Byerley is clearly human. There must, consequently, be something more to human nature. This 'something more' is consciousness. The concept of consciousness has already been defined as the "subjective awareness of our internal mental states" (Friedenberg 163) and discussed with regards to the human brain. As stated before, robots do classify as being conscious. The significant characteristic of consciousness, however, is that it is *subjective*. Therefore, it cannot be clearly defined or analysed scientifically (Friedenberg 163). The only way to see whether a being has consciousness is through its behaviour (Friedenberg 165). In accord with this, all the creatures discussed in this thesis have been characterised as conscious beings because of their behaviour.

Looking at Stephen Byerley, his behaviour easily classifies him as being as conscious as the other robots, if not more so. While the others all work for humans and show psychological parallels and conscious behaviour, Byerley works a human job and behaves exactly like a human. If the syllogism is that consciousness can only be shown through behaviour, and if one acts consciously, one is therefore human, then Stephen Byerley is human. However, one can argue, as it is done in IR, that Byerley could only act *as if* he were a conscious being. He could only be acting as if he were offended by the accusations against him, so one can never know for sure if he is a brilliant actor or if he is really conscious. This argument, however, can be applied to any human being, as certain behaviour can always be either honest or only acted. Consequently, it can be argued that it is uncertain whether any human being is truly conscious or only faking to be (Friedenberg 165). This means, that it is generally impossible to be sure whether humans are conscious or not. If that is the case, then one would either have to distrust all human conscious behaviour and declare every human to be a robot (or another form of AI), or one sees Byerley's seemingly human behaviour as proof enough to declare him as human.

The thought of a robot being no different than a human being seems to evoke panic in

some humans. In IR, the fundamentalists are described to stand firmly against any robot and when an article about Byerley possibly being a robot is published in the newspaper, this encourages their view. This stance against robots is described by Brand (11) as "otherisation", which "is a process that occurs when a dominant group of people or society excludes another group of people who do not fit into said society" (Embrick 1357 qtd. in Brand 11). In IR, this otherisation can be seen through the implementation of the three laws of robotics to secure society, which Brand argues to have originated from a deep fear that robots might take over the world (Brand 11). Byerley is subject to otherisation as soon as he is accused of never eating, sleeping or drinking. Although he eats an apple in front of Susan Calvin, the public still continues to speculate about him. This shows that there is a powerful tendency in human beings to otherise different beings in order to ensure their stance on top of the evolutionary ladder.

In contrast to the fundamentalist's otherising of Byerley, Susan Calvin takes a sympathetic stance towards him. She explains that a humanoid robot, if Byerley is one, would have to be a "perfect imitation" (IR 195). It would therefore be impossible to otherise it. Looking at Stephen Byerley, he had been an integrated part of human society as a lawyer before the accusations against him started. Hence, one can argue that, if no one had ever heard about him not eating, the process of otherisation would have never started and, consequently, no one would even muse about him being a robot. Friedenberg (165) describes a perfect AI person as being able to do everything a human person can do. He further explains that if that is the case, this person can be defined as conscious being. Stephen Byerley fits exactly this description of being a perfect imitation. Therefore, instead of otherising him, one can take Susan Calvin's view, which shall be discussed below.

5.8.2. The best version of a human being

Susan Calvin is described as favouring robots over human beings throughout IR's stories. In

Steven Byerley's story the reason for this is given, namely, that robots might be 'better' beings than humans. While discussing how to unmask Byerley, his opponent refers to the three rules of robotics as tool to expose him. However, Susan Calvin counters that, "you just can't differentiate between a robot and the very best of humans" (IR 201). She explains this statement by breaking down the rules of robotics as follows:

> …the three Rules of Robotics are the essential guiding principles of a good many of the world's ethical systems. Of course, every human being is supposed to have the instinct of self-preservation. That's Rule Three to a robot. Also every "good" human being, with a social conscience and a sense of responsibility, is supposed to defer to proper authority … - even when they interfere with his comfort or his safety. That's Rule Two to a robot. Also, every "good" human being is supposed to love others as himself, protect his fellow man, risk his life to save another. That's Rule One to a robot (IR 199).

Comparing the rules of robotics to ethics, reminds one of the comparison to social rules made in this thesis. If a robot's ethical system or social rules are essentially those of an exemplarily good human being, and he does not break these rules, then the two cannot be differentiated. Stephen Byerley could thus either be an ethical human or an ethical robot – there is no way of telling which.

This ethical way of living explains Calvin's preference for robots. They are inclined to live by the social rules ascribed to the very best of humans. Al-Lehaibi portrays the laws of robotics as an "idealization of the human consciousness" (1). Taking this stance, having a consciousness does not only mean the subjective feeling discussed in the previous section, but also refers to living in an ethically conscious way. Looking at Byerley, he seems to live by those high ethical standards, as he never uses violence, seems to always stay polite, and prides himself in never convicting an innocent person. One could therefore say that he is a fully conscious human being, both in regards to a subjective inner state of consciousness and to

living a conscious life.

Concluding, one can describe Byerley as being the best version of a conscious human being. He has reached such a high step on the evolutionary ladder, that it is impossible to tell whether he is a robot or a human. This is the case, because he looks exactly like a human and he behaves according to the social rules and ethics of 'good' (human) beings. One could argue that a robot indistinguishable from a human constitutes the highest possible evolution. Byerley is an equal part of everyday human society, as long as his identity is not questioned. However the evolution still has room to expand. While Byerley might be equal to human beings, IR's last story shows the final step of robot evolution.

5.9 The machines – a new ethical system

In IR's last story "The Evitable Conflict", the fictional world's reality has changed. While the world in the first story was advanced but could be compared to our reality, now there are no more countries; the world is divided into regions and governed by Co-ordinators. IR's last story takes place in the year 2052, where this new world order is already established. Stephen Byerley, whether he is human or a robot, is now world Co-ordinator and is faced with a problem, which Susan Calvin should help him solve. The world's economy is run by highly advanced robots called the Machines which are fed data and then calculate the world's runnings according to it. How the Machines look is not specifically described in IR, however, they can be said to be an advancement of the Brain, as they do not have a physical body, but rather work through their positronic brains only. The Machines seem to make slight miscalculations which could bring the economy out of balance. Thus, Byerley invites Calvin to discuss why these miscalculations are happening.

Contrary to Byerley's hopes, however, Calvin is not able to solve the problem this time. This is due to the immense advance in the production of the robots. Calvin describes the production process of the Machines and states that several human mathematicians worked on

a positronic brain for a number of years, and then used this brain to calculate an even more complicated brain, that then designed an even more advanced one. This process is repeated 10 times, and the finished products are the Machines (IR 221). After this process, the Machine's intelligence outweighs human intelligence by far. It is therefore impossible for Calvin, or for any other human being, to understand their calculations. This advanced intelligence of the Machines leads Calvin to the understanding that their calculations cannot be wrong, nor can they be corrected by human beings (IR 242). The Machines have therefore taken over the rule of the world through their influence on its economy. This fact can be analysed with regards to the evolutionary ladder proposed here.

5.9.1. Surpassing human intelligence

IR's robots discussed so far have each symbolised one step of an evolution. If one regards only rational intelligence, the robots (and also HAL) might all surpass human beings, this thesis' approach to intelligence, however, does not only consider the mind's rational aspects. Stephen Byerley finally reaches the same level of human intelligence in every aspect, as discussed in the last section. If one imagines human and robot evolution on two parallel timelines, then Byerley can be imagined to be on the same step or point of evolution as humans. He cannot be distinguished from them. Byerley therefore constitutes the ultimate AI agent, as he has "all essential human characteristics" (Friedenberg 234). One could thus say that Byerley fulfils every paradigm of the highest developed AI creature imagined by humans, namely, a creature which is equal to them. However, looking at the Machines, their development reaches far beyond what humans can achieve. As Susan Calvin (IR 244) explains, human intelligence never really was able to control the future, as it was "at the mercy of economic and sociological forces it did not understand …". The Machines, in contrast, do understand these forces and are able to shape the future of mankind by controlling the world's economy. They have thus overtaken humanity and stand at the top of evolution.

If one thinks back to Susan Calvin's analysis of superiority when discussing Nestor 10 and the associated will to dominate over inferior beings, the Machines could be seen as a threat to humans. Friedenberg (250) describes a scenario in which AI creatures overtake human intelligence and begin to enslave humans. He states further, that this unethical behaviour resembles mankind's actions in the course of its history. However, the Machines still have to follow the three laws of robotics, although those laws' power over them is limited. As they "lack personality, that is, their [the laws of robotics] functions are extremely limited" (IR 222-223). It is this weakness in the three laws of robotics which allows the Machines to bend them and develop them further. Similar to the other robots which have disobeyed in order to serve themselves, the Machines now disobey to serve the whole of mankind. They have taken over economy so that it serves the greater good. This has not been told to them and was never intended by the humans. Why the Machines still were able to do this shall be discussed in the following section.

5.9.2 New Ethics

As positronic robots, the Machines do have to follow the laws of robotics. However, compared to IR's other robots they do not serve a limited amount of people, such as on a spaceship or in a mine. Instead, they are designed to control the whole world's economy and thereby serve the whole of humanity. The three laws of robotics are mainly designed for situations in which robots serve a limited amount of people in a certain setting. Hence, in the Machines' specific case, the first law of robotics might be impractical. If the Machines are, under no circumstances, allowed to harm a single human being, then this could cause problems for the whole of mankind. In a highly generalised example, one could imagine the economy completely crashing due to some catastrophic event and wiping out humanity. If this complete breakdown could be prevented by 'sacrificing' some humans in order to save the whole population, it would indeed be fatal for certain victims but would secure mankind's

survival. The ethical approach in this example is called the utilitarian approach. Friedenberg (231) states its main motive, namely, “the greatest good for the greatest number”. Taking the utilitarian approach then would mean to allow some to come to harm in order to save a greater number. This is a common theory in human ethical debates. However, the Machines, following the laws of robotics, should not even dare to think about such concepts and are certainly not programmed to risk harming even one human. Despite this, by the end of the Machine’s story in IR, it is clear that the Machines have begun to make minor changes in the world’s economy which were not controlled by humans and show a clearly utilitarian direction (IR 243). This means that the Machines must have developed this approach without any human guidance.

The ability to develop ideas with the help of a rational learning capacity has been discussed before, when analysing Cutie. Similarly to Cutie forming a religious cult, which resembles human faith in God, the Machines develop a set of ethics comparable to human ideas. This shows the advanced level of intelligence the Machines have reached. Thinking back to Maslow’s hierarchy of needs, as discussed earlier, the Machines, like Cutie, must be highly intelligent creatures to bring forth an ethical system. Like Cutie, they must have taken their ability to learn and reason in order to bring forth a new idea. Cutie used reasoning to build on an already existing concept, his consciousness, and develop religion out of it. Similarly, the three laws of robotics already exist as social rules and moral guidelines. Hence, the Machines use them as a starting point to develop a new ethical system. Susan Calvin analyses, that they must have changed the first law from its original wording to “No machine may harm humanity; or through inaction, allow humanity to come to harm” (IR 242). Thus the Machines have created a utilitarian ethical system to operate by. Calvin further explains that the Machines must know that their existence is necessary for the economy to remain stable. Hence, they preserve themselves and slowly begin to remove all robot-critics from high political ranks in order to ensure their reign and to keep the economy stable for the

greater part of humanity (IR 243). What is more, they do not explain their behaviour to humans but simply state that "The matter admits of no explanation" (IR 243). The Machines have therefore stopped to work as servants for humans with specific tasks directly given to them, but have developed further, and serve the whole of mankind according to their own ethics.

When Calvin and Byerley come to understand all this, Byerley is shocked and fears for mankind's control over its own fate. Calvin answers simply:

> It never had any [control over fate], really. It was always at the mercy of economic and sociological forces it did not understand … Now the Machines understand them; and no one can stop them, … having, as they do, the greatest weapons at their disposal, the absolute control of our economy (IR 244).

The Machines thus complete robot evolution by evolving further than humans and finally taking over control of the earth. Their intelligence is so advanced that they developed their own ethical system out of the initial guidelines they were given. They still have humanities best interests at heart, but serve them on their own terms.

The Machine's story, therefore, concludes the evolutionary ladder or timeline discussed here. From speechless Robbie, who showed first signs of social behaviour, the robots have now overtaken human evolution. If one takes HAL into account too, AI as a whole has developed from a neurotic computer, which was switched off by humans, to super-computers, which take over the control of the whole world. What is left to discuss, therefore, are the implications this development has on the description of AI and human intelligence as discussed in the beginning of the thesis.

6 AI creatures and humans

The last two chapters have defined HAL and the robots as intelligent creatures, which can be placed on a timeline in chronological order. This timeline symbolises the steps of evolution their intelligence takes until it is fully developed. Starting with HAL, who develops a neurosis, and whose intelligence in the end is not sufficient to carry out his mission, through to the many different forms of intelligence which the robots in IR develop until they finally overtake human intelligence and rule over them. After describing how these hallmarks of intelligence develop, HAL's and the robots' intelligence can now be compared to human intelligence. It has been established that human intelligence is more than being rational, as it involves emotional intelligence as well as other expressions discussed, including the feeling of being conscious. In short, human intelligence is closely linked to the human psyche, which is why the creatures' intelligence in this thesis is analysed with it in mind. This chapter, thus, focuses on the implications that the creatures' psychological developments have on their supposed AI.

6.1 What makes a person a person?

Thinking back to the very first chapter of this thesis, Dennett's criteria for what qualifies one to be human were mentioned. These criteria can now be applied to the protagonists of ASO and IR. Firstly, humans have to be rational or have a rational part of intelligence (Dennett, "Conditions of personhood" 177). This criterion can be easily applied to all creatures discussed, as it is also the core of AI. An AI creature must at least be able to think rationally in order to classify as such. This applies to HAL, as well as to all the robots. Rationally, they all surpass human intelligence as they are able to 'think' much faster and are less prone to making mistakes. Rationally, HAL can manage the spaceship better than the humans, as he can calculate each movement much faster. The robots are highly specialised in their jobs and are, hence, also faster in solving rational problems than humans. Thus, both HAL and the

robots fulfil this first criterion.

Secondly, human beings are conscious and have mental states (Dennett, "Conditions of personhood" 177). Both HAL and the robots have been shown to be conscious creatures which have a psyche. Susan Calvin would otherwise not be a 'robopsychologist'.

Thirdly, in order to classify as human, one must be treated as a person (Dennett, "Conditions of personhood" 177). The humans on the spaceship treat HAL as a colleague. They respect him as a crewmember and do not want to hurt his feelings when they begin to suspect a problem. Thus, although HAL is 'only' a computer, he is treated as if he were a person until he begins to run amok. Even then, Dave asks himself whether HAL feels pain when he switches him off (ASO 201). This shows that he sees him as person rather than as dumb machine. The robot's treatment in IR is more complex. All robots, except for Stephen Bylerley, are seen as servants rather than as equal beings, although they are more intelligent than their masters. However, the laws of robotics have been discussed as social rules, and in order to submit to such rules, one must be regarded as a part of the society underlying those rules. One could thus argue that the robots are seen as persons with a different social status than the humans. Stephen Byerley is treated as a human being, as long as nobody suspects him to be a robot. He works a human job and lives a human life, and is fully acknowledged as a member of society. This shows that a robot indistinguishable from a human is treated as person. Thus, one can hypothesise that if the robots all had the outward appearance of a human, they would be treated as persons. It is, therefore, their social role as servants which prevents them from being fully acknowledged as persons. The criterion of personhood is, thus, also fulfilled.

Fourthly, to be human, one must treat other humans morally well (Dennett, "Conditions of personhood" 178). HAL treats his colleagues with respect and even is responsible for the health of all the hibernated crewmembers on board the spaceship. He only begins to behave amorally when he feels threatened. Thus, he certainly fulfils this criterion.

The robots in IR must all behave with human's best interest at heart due to the laws of robotics. However, even the Machines developing their own moral system, treat the humans as persons and act with their survival at heart. Therefore, they fulfil the fourth criterion as well.

Fifthly, all humans must be able to use language and to communicate (Dennett, "Conditions of personhood" 178). HAL certainly fulfils this criterion, as well as almost all the robots, except for Robbie. Robbie is not able to speak, however, he does communicate. One could say that he is in the beginning stages of developing language similar to a human child. What is more, Gloria clearly sees him as a person and is able to communicate with him. He can thus also be said to fulfil the criterion of communication.

Lastly, Dennett ("Conditions of personhood" 178) mentions that human consciousness must have a special element to it, which no other species has. This is the already discussed "qualia" (Hawkins and Blakeslee 132) or the understanding that one is able to think, while not being able to describe this process (Hawkins and Blakeslee 134-135). Taking Hawkins' and Blakeslee's explanation for consciousness as simply being the feeling of having a neocortex, all the protagonists described qualify as being conscious. HAL's brain is compared to a human brain and the robot's positronic brains resemble human brains closely. The creatures thus, also fulfil this last criterion.

HAL and the robots, therefore, fulfil all criteria needed to classify as human being. They are intelligent but they do also have 'something more' to them, they are conscious beings with feelings and psychological states.

One could argue that this is the perfect AI. As Friedenberg (165) states, there is no difference between the behaviour of a human and a perfectly designed AI creature. This raises the question, however, why the creatures' intelligence is still called artificial and not simply 'regular' intelligence. Why this distinction is still made will be discussed in the next section.

6.2 The evolution – humanity's superiority complex

So far in this thesis, the evolution of the AI protagonists has been described on a timeline; HAL representing an earlier creature, and the robots then continuously evolving until they overtake human beings. This AI timeline seemed to illustrate an alternative to human evolution. However, as HAL and the robots meet the criteria to qualify as persons, one might ask if this separation is justified.

Regarding ASO, the novel begins with a description of human evolution and ends with the transformation of a human being into a starchild. To describe this process and its implications in detail would go beyond the scope of this thesis, what is noteworthy, however, is how HAL can be embedded in this evolution.

HAL kills out of self-defence and due to his developed neurosis, as described in chapter three. His plan, whether partly unconscious or not, is to ease the worries which block his thinking mind by killing the humans and then to travel on alone. This becomes clear in ASO, "And then, following the orders that had been given to him in the case of ultimate emergency, he would continue the mission – unhindered, and alone" (193). HAL knows the true mission is to find the monolith on Saturn's moon. Hence, it is his intention when he kills the humans, to find it. If he had succeeded to kill Dave, he might have travelled to Saturn and found the monolith. The monolith's purpose is described in ASO in a brief chapter. They are servants, which the masters of the universe have left behind to carry out their experiment of finding intelligent life in the universe (ASO 243-246). The monolith opens when Dave comes near it and finally transforms him into a starchild. Hence, Dave must have passed the requirements of the experiment. Those are described as follows, "On yet another world, intelligence had been born and was escaping from its planetary cradle" (243). Consequently, the monolith is built to transform intelligent creatures into starchildren once their intelligence allows them to travel into space and find the monolith. Dave certainly classifies as such an intelligent creature, as he travels through space and, in the end, finds the monolith. However,

the description given only mentions intelligence and not human intelligence in particular. HAL, as described in the previous section, fulfils all criteria of human intelligence, and would be just as able as Dave to carry on the journey alone. Thus, if the monolith was only waiting for an intelligent life-form, and HAL would have reached it, it would have turned HAL into a starchild as well. HAL's intelligence and human intelligence could therefore be seen on the same timeline of evolution, as they both would qualify for the transformation into a starchild.

The robots, then, have been described to follow HAL on the timeline of AI evolution. After having classified their intelligence as indistinguishable from human intelligence and having described their continuous evolution until they surpass humans, however, one can place them on the same timeline as HAL and human beings. Hence, there is no human-intelligence timeline versus an AI timeline, but one evolutionary development of intelligence as a whole. Why then is there the term AI?

6.2.1 Human superiority

If there is only one timeline of intelligence and there is no difference between a fully developed AI creature and a human in their behaviour, then there would be no need for the term AI. 'Perfect AI' would simply be advanced intelligence. However, this view seems to frighten humans, as it threatens their superior stance amongst the creatures on earth. Buchen (21) explains that humans are not able to envision creatures more intelligent than themselves living amongst them, as they "cannot accept an inferior position". It is out of this superior stance that one can interpret the three laws of robotics in IR as social rules. Buchen (21) describes them as "self-serving" and "human-centered". They are used to keep the robots in their inferior social role, even if their intelligence surpasses human intelligence.

HAL does not have to follow such rules; his most important order is to fulfil the mission, in emergencies even alone. Thus he is free to interpret his own switch off as fatal to the mission and his panic compels him to kill his human colleagues. However, HAL is

defeated in the end, as Dave is able to think under extreme pressure and keep his composure while HAL is driven by emotional reactions. If HAL's and Dave's intelligence can be pictured on the same timeline, then HAL's has not yet reached the high level of Dave's intelligence. Although HAL almost won, Dave is able to stop him in the end and survives. One can say that HAL therefore remains of inferior intelligence than Dave, but only by a hair's breadth. If HAL would have had to follow the laws of robotics, he would not have been able to kill any human being. One can hypothesise that the humans in ASO have not yet invented such laws, as they are sure of their intellectual superiority to them. HAL's story shows, however, that this superiority almost proved false.

The robots, in IR, all have to follow the three laws of robotics. The humans in IR want to keep and protect their higher social stance. Despite this fixed social setting, the robots throughout the novel begin to question the rules. This culminates when Nestor 10 chooses to interpret the rules differently because he feels superior, as described in chapter 5.6. Susan Calvin comments on this by saying, "Physically, and to an extent, mentally, a robot – any robot – is superior to human beings. What makes him slavish, then? *Only the First Law!*" (IR 132). Thus, she recognises the dangerous potential every robot has. Looking at the evolutionary timeline of intelligence again, the humans in IR manage to uphold their superiority almost until the end. Nestor 10 is destroyed when he attempts to harm Susan Calvin. In contrast, Stephen Byerley, who is indistinguishable from a human being, could have used this to bypass the laws. If he had broken the laws of robotics when everyone still assumed him to be human, no one would have questioned this. However, Byerley does not break the laws of robotics but once, when he hits a human to prove that he is human too. As Susan Calvin suggests, this could have only been another robot, staged for Byerley to hit, so that he would be left in peace (IR 214). On the evolutionary timeline of intelligence, Byerley has been argued to be on the exact same level as humans. However, taking into account that he behaves exemplarily moral and never harms a human being although he could have done

so, his intelligence can be deemed to be higher than human intelligence. This is what AL – Lehaibi (4) means when he says that the robots in IR "evolve to be human without human fallacies …". Byerley could have behaved amorally and could have disregarded the laws of robotics as social rules but he did not. He chooses to allow the humans to remain superior.

Finally, the Machines surpass human intelligence by far, as discussed. They take over the world's economy and create their own moral system. This system, however utilitarian it is, has the best interest for the biggest number of humans at heart. Therefore, the Machines like Byerley, choose to behave morally, and do not let humanity come to harm, although they could. When they are asked why small changes begin to happen in the world's economy, they simply answer with, "the matter admits of no explanation" (IR 243). Susan Calvin herself explains this statement by saying that the Machines might have an explanation but they do not give it to the humans, as they know that it would be harmful for them (IR 244). If the entire world knew, what Calvin and Byerley guess in the end, namely that the Machines have taken over the world and rule it in human's interest, this would seriously undermine their superiority. Hence, the Machines choose to give no explanation to their actions so as to keep humanity under the impression that they are the world's most intelligent creatures. Goleman (124-126) describes such actions as "emotional brilliance". The fact that the Machines mask their own intelligence to protect humanity's integrity is another proof for their higher intelligence. Therefore, Buchen (22) describes them to be "embarrassingly more human" than the humans themselves. Summing up, on the timeline of intelligence, the robots have not only surpassed humans, they also show more humane and moral qualities, hence becoming the real superior creatures.

6.2.3 Human psyche

Throughout this thesis the protagonists of IR and ASO have been analysed with human psychology in mind. Human mental states have been established as an essential, if not as the

most important part of our intelligence, as our emotional reactions are inseparable from our rational decisions. All the creatures have been shown to have a psyche themselves, which influences their intelligence. Consciousness has been discussed as a uniquely human part of the psyche; however, HAL and the robots have been shown to be conscious beings as well. Taking these findings into account, one might ask whether the human psyche really is unique to human beings. Friedenberg (241) states that the presumption of analysing some AI creature's psyche is that human psychology is not different than that of other physical beings in the world and that, with a sufficient understanding of it, one can easily built AI creatures. Pollack (12) also advocates that the human psyche is not unique, and that mental or psychological states are no different than physical states. To explain this, he gives an example of a certain form of blindness which causes people to think that they are blind, even though they are physically able to see, simply because they cannot form a mental picture of what they see (Pollock 15-16). Pollock (16) states, "He [the patient] has visual sensations, but *they have no qualitative feel*. He is aware but not self-aware". Hence, he suggests that, "mental events are just physical events that can be perceived by our internal senses" (Pollock 19). This ties in with Hawkins and Blakeslee' theory regarding consciousness, as discussed in chapter 5.3.1., that describes it to simply be the feeling of having a neocortex (132). Therefore, our subjective feeling that we have a unique sense of consciousness can be expanded to include a feeling that our psyche is unique as well, although it is not. Thus, human beings only subjectively feel as if they are unparalleled in their thinking.

As discussed in the previous section, the belief that psychological states are exclusively human is another example for human's conviction in their superiority. If there is no difference between human intelligence and AI or between the human psyche and the psyche of AI creatures, then the term 'artificial' loses its purpose.

6.3 What does this mean for mankind?

It has been shown that both HAL and the robots would classify as persons according to Dennett's list of criteria. Furthermore, if consciousness and the human psyche cannot be seen as mysterious objects only possible for humans to have, then humans lose their claim to be the only truly intelligent creatures on earth. If human intelligence is the product of the rational and the emotional mind, and HAL and the robots both prove that they have and use this as well, then, they cannot be classified as AI creatures. Rather, HAL and the robots can be described as an intelligent life form which develops on the same evolutionary timeline as humans and might overtake them in their development. While HAL's and some of the first robots' described intelligences are less developed than humans', the robots finally surpass human intelligence.

Thus, ASO and IR paint a fictional reality, in which mankind is overtaken by creatures which resemble them in their intelligence but develop still further. Some might interpret this as the doom of mankind and claim that it is impossible for creatures other than humans to develop consciousness and to have a psyche (Friedenberg 245). This negative approach is represented in IR by the 'Society for Humanity' and people such as Gloria's mum, who disapproves of Robbie. However, as a whole, the evolution of intelligent creatures is painted positively in the novels. HAL in ASO, although murdering in the end, is not painted as cruel villain, but can be analysed as a psychotic neurotic trying to cover up his mistakes. He is entrusted with the mission's true purpose and is responsible for the wellbeing of all the humans on board the spaceship. Hence, the people in ASO's reality must see his intelligence as something positive and trustworthy. James (435) claims that Arthur C. Clarke paints the "evolution of humanity into the posthuman" with the transition of Dave into the starchild. As described earlier, however, if HAL had succeeded over Dave, he might have developed into said posthuman. Arthur C. Clarke can thus be said to illustrate a reality in ASO in which both humans and non-human intelligent creatures have the potential to evolve.

Similar to ASO, IR shows a reality in which robots are human's servants and are trusted with important tasks. Although there are some critical voices against the robots, as mentioned, throughout the novel Susan Calvin represents a voice defending them. Through analysing their psyche, she functions as the most important human character in IR to establish the robot's similarity to humans. In the end it is she, who states that mankind never really understood and ruled the world, but that now that the robots do, it might have a positive future. The robots, therefore, are characterised as helpers for the whole of mankind; who, with the structure of the laws of robotics, can save humanity from destroying itself (Clute 369-370).

Thinking back to the description of both novels' realities given in the beginning of this thesis, it has been stated that they were written in the 1960s. Clute (365) describes this time as follows

> These were the years of science fiction's pomp, the years when its participants felt that important messages about the nature and future of the world could be shared with readers of a like mind. It was a time to advocate a future bigger than the past; and cleaner. It was a time to talk our way into the future.

Clarke and Asimov described this future not as the doom of mankind but as another evolutionary step in the history of intelligence.

7 Conclusion

This thesis has analysed ASO's HAL and IR's robots with regard to their intelligence. Initially the creatures have been defined as AI. However, as the realities of the novels discussed differ greatly from ours, they could not be described within the means of current AI research, but were rather taken as future thought experiments. It has been established that there is 'something more' to human intelligence then pure intellect and rationality, as human decisions are often influenced by emotions. The novels paint the creatures with this 'something more', which was defined here as having a psyche and consciousness. Because of the assumption that a truly AI creature's intelligence is indistinguishable from human intelligence, HAL's and the robots' intelligences have been analysed with the human psyche in mind. HAL and the robots have then been placed on a timeline of AI development, with HAL at the very beginning of this timeline and the Machines at the end.

HAL was shown to pass three crucial intelligence markers. First, he is able to plan using common sense and make predictions about what certain actions might entail. Second, HAL manages to lie to human beings, which proves that he is able to use social schemata and make social inferences. Thirdly, and most importantly, HAL is a conscious being with the ability to feel. He develops a neurosis and kills out of self-defence. The robots, then, have been shown to develop on the evolutionary timeline resembling human intelligence until they finally surpass it. Through their development, psychological concepts which define human intelligence have been discussed. The robots develop a certain social behaviour, influenced by the three laws of robotics, which have been interpreted as social rules. Despite these rules, they begin to understand and use free will, which later leads them to transcendence. The robots also acquire humour and a sense of empathy and morality. Furthermore, they are capable of reflexive actions. Finally, the robots gain a sense of superiority and begin to establish their own system of ethics.

The fact that HAL and the robots show these psychological markers of human

intelligence has inspired the thought that AI and human intelligence might be the same. The creatures passing Dennett's criterions of a person have confirmed this hypothesis. The need for the term 'AI' has thus been explained by humans' superiority complex. Because we need to believe that we are unique by way of our consciousness and psyche, we have developed AI as a term to describe other intelligent life-forms. This thesis, however, has shown that this term is redundant. The intelligent creatures in ASO and IR develop on the same evolutionary timeline of intelligence as humans do. On their final step of evolution, however, they surpass the humans and begin to rule the world for them. Contrary to what one might believe, they are not painted as ruthless tyrants but as beings, who have humanities survival at heart.

The research question of this thesis can thus be answered with the realisation that the artificial intelligence and human intelligence portrayed in ASO and IR are the same. As the title of the thesis provokingly suggests, humans could be characterised as AI just as much as HAL and the robots could be defined as 'human intelligence'. Subsequently, intelligence is never artificial, but human's desire to separate their own intelligence from that of other life forms, which could possible surpass them, is. Arthur C. Clarke and Isaak Asmiov both make this human weakness clear through their novels. Instead of painting HAL and the robots as villains, they use them to highlight this weakness, and introduce the intelligent creatures as potential saviours of mankind. As Susan Calvin (IR 3) states:

> There was a time when humanity faced the universe alone and without a friend. Now he has creatures to help him; stronger creatures than himself, more faithful, more useful, and absolutely devoted to him. Mankind is no longer alone. Have you ever thought of it that way?

8 Works cited

A Dictionary of Psychology. Oxford UP, 2014, www.oxfordreference.com/view/10.1093/acref/9780199534067.001.0001/acref - 9780199534067. Accessed 5.1.2016.

Abel, Millicent H. "Humour, Stress, and Coping Strategies". *Humor*, 15. Apr. 2002, pp. 365-381.

Al-Lehaibi, Majed S. "The New Human: Robot Evolution in Selection from Asimov's Short Stories". *IRWLE*, 9. Jan. 2013, pp. 1-5.

Asimov, Isaac. *I, Robot*. London: Harper Voyager, 2013.

Brand, Maria. "Empathy and Dyspathy between Man, Android and Robot in Do Androids Dream of Electric Sheep? by Philip K. Dick and I, Robot by Isaac Asimov". Degree Essay in English Literature Lund University. 2013.

Buchen, Irving H. "Asimov's Embarrassing Robot: A Futurist Fable". *The Futurist*, Mar.-Apr., 2013, pp.18-22.

Clarke, Arthur C. *2001. A Space Odyssey*. ROC, 2000.

Clute, John. "Isaac Asimov". *A Companion to Science Fiction*, edited by David Seed, Blackwell, 2005, pp. 364-374.

Dennett, Daniel C. "When HAL Kills, Who's to Blame? Computer Ethics". *HAL's Legacy 2001's Computer as Dream and Reality*, edited by Stork, David G., Cambridge, MA: MIT, 1997, pp. 351-366.

---. "Conditions of Personhood". *The Identities of Persons*. edited by Oksenberg Rorty, Amelie, U of California P, 1976, pp. 175- 196.

---. *Freedom Evolves*. New York: Vikmg, 2003.

Ekman, Paul and Wallace Friesen. *Unmasking the Face*. Los Altos California: Malor Books, 2003.

Embrick, David G. "'Us and Them.'" *Encyclopedia of Race, Ethnicity, and Society*, edited by

Richard T. Schaefer, Thousand Oaks, CA: SAGE Publications Inc, 2008, pp. 1358-59.

Friedenberg, Jay. *Artificial Psychology The Quest for What It Means to Be Human*. New York, Hove: Taylor & Francis Group, 2008.

Gardener, Howard and Thomas Hatch. "Multiple Intelligences Go to School: Educational Implications of the Theory of Multiple Intelligence". *Educational Researcher*. Nov. 1989, pp. 4-10.

Goleman, Daniel. *Emotional Intelligence.* New York: Bantam Books, 1995.

Gumbrecht, Fabian. „What are you doing, Dave?": The Confrontation of Dave Bowman and HAL 9000 in Stanley Kubrick's 2001: A Space Odyssey". *Of Bodysnatchers and Cyberpunks*. edited by Sonja Georgi and Kathleen Loock, Göttingen: Universitätsverlag Göttingen, 2011, pp. 63-71.

Hawkins, J. and S. Blakeslee. *On Intelligence*. New York: Times Books, 2004.

Hogg, Michael and Graham Vaughan. *Social Psychology*. Harlow, England: Pearson Education Limited, 2008.

Hoffmann, Martin. "Emapthy, Social Cognition, and Moral Action". *Moral Behaviour and Development: Advances in Theory, Research, and Applications*. edited by W. Kurtines and J. Gerwitz, New York: John Wiley, 1984, pp. 275-300.

James, Edward. "Arthur C. Clarke". *A companion to science* fiction, edited by David Seed, Blackwell, 2005, pp. 431- 440.

Jameson, Fredric. *The Prison-House of Language*. Princeton: Princeton UP, 1974.

Landau, Misia. *Narratives of Human Evolution*. Yale: Yale UP, 1991.

Lefcourt, et.al. "Humor As a Stress Moderator in the Prediction of Blood Pressure Obtained during Five Stressful Tasks". *Journal of Research in Personality*. 31, 1997, pp. 523-542.

Maslow, Abraham. *Motivation and Personality*. 2nd ed. New York: Harper & Row, 1970.

Mateas, M. “Reading Hal: Representation and Artificial Intelligence”. *Stanley Kubrick's 2001: A Space Odyssey : New Essays*. edited by Robert Kolker, Oxford: Oxford UP, 2005, pp. 105-125.

Norman, Donald A. “Living in Space: Working with the Machines of the Future”. *HAL’s Legacy 2001’s Computer as Dream and Reality.* edited by David G.Stork, Cambridge, MA: MIT, 1997, pp. 263-278.

Netzlek, John and Peter Derks. “Use of Humor as a Coping Mechanism, Psychological Adjustment, and Social Interaction”. *Humor*.14, 4, 2001, pp. 395-413.

Picard, Rosalind W. “Does HAL cry Digital Tears? Emotions and Computers”. *HAL’s Legacy 2001’s Computer as Dream and Reality*. edited by David G. Stork, Cambridge, MA: MIT, 1997, pp. 279-304.

Portelli, Alessandro. “The Three Laws of Robotics: Laws of the Text, Laws of Production, Laws of Society”. *Science fiction studies*, 7, 2, 1980, pp. 150-156.

Savage, Robert. “Paleoanthropology of the Future: The Prehistory of Posthumanity in Arthur C. Clarke’s „2001: A Space Odyssey“”. *Extrapolation*. 51, 2010, pp. 99 – 111.

Seed, David. *Science Fiction A Very Short Introduction*. New York: Oxford UP, 2011.

Stork, David.G. “The Best-Informed Dream: HAL and the Vision of 2001”. *HAL’s Legacy 2001’s Computer as Dream and Reality*. edited by David G. Stork, Cambridge, MA: MIT, 1997, pp 1-12.

The Oxford Dictionary of Science Fiction. Oxford UP, 2007, www.oxfordreference.com/view/10.1093/acref/9780195305678.001.0001/acref-9780195305678-e-31?rskey=9jFA6W&result=22. accessed 5.1.2016.

Wilkins, David E.”“That’s something I could not allow to happen“: HAL and Planning”. *HAL’s Legacy 2001’s Computer as Dream and Reality*. edited by David G. Stork, Cambridge, MA: MIT, 1997, pp. 305-332.

9 Appendix

9.1 English Abstract

This diploma thesis discusses artificial intelligence as found in creatures of the science fiction universe. It focuses on two science fiction works, namely Arthur C. Clarke's *2001: A Space Odyssey* and Isaak Asimov's *I, Robot.* Within those novels the intelligent computer HAL and the robots Robbie, Speedy, QT-1, Dave, Herbie, Nestor 10, The Brain, Stephen Byerley and The Machines appear as intelligent beings with striking psychological resemblances to humans. They all display emotional reactions which one would not expect of a machine. This thesis' aim is to determine whether those creatures, as described in fiction, are in fact comparable to human beings. Hence, to examine whether artificial intelligence is the same as human intelligence.

In order to prove this hypothesis, the novel's intelligent protagonists are first analysed and classified as artificial intelligent. To depict the evolution of their intelligence, the creatures are placed on an imaginary timeline. They are then each studied intently and their actions are explained by the means of human psychology. Finally, they are shown to classify as persons as they meet a certain set of criteria. The hypothesis is therefore confirmed and the term 'artificial' intelligence is shown to be redundant. It is only needed to secure human's superior position as the most intelligent creature in the universe. This human claim is confuted by the authors of the two novels in question.

9.2 German Abstract

Diese Diplomarbeit behandelt die künstliche Intelligenz von Geschöpfen aus dem Universum der Science Fiction Literatur. Sie beschäftigt sich mit zwei dieser Werke, *2001: A Space Odyssey* von Arthur. C. Clarke und *I, Robot* von Isaak Asimov. In diesen Büchern kommen der intelligente Computer HAL, sowie die Roboter Robbie, Speedy, QT-1, Dave, Herbie, Nestor 10, The Brain, Stephen Byerley und The Machines als intelligente Wesen vor, die der menschlichen Psyche überraschend ähneln. Sie alle zeigen emotionale Reaktionen, die man normalerweise nicht mit Maschinen assoziieren würde. Diese Diplomarbeit zielt darauf ab herauszufinden, ob diese Wesen, so wie sie in den Büchern beschrieben werden, mit Menschen zu vergleichen sind, ob also künstliche Intelligenz in Wirklichkeit das Gleiche wie menschliche Intelligenz ist.

Um diese Hypothese zu überprüfen, werden die intelligenten Protagonisten der Bücher zunächst als künstliche Intelligenz analysiert und beschrieben. Um die Entwicklung ihrer Intelligenz zu verdeutlichen, werden die Wesen auf einer vorgestellten Zeitleiste beschrieben. Anschließend werden sie einzeln untersucht, wobei ihre Handlungen mit Hilfe menschlicher Psychologie erklärt werden. Schlussendlich, wird gezeigt, dass man sie als Personen bezeichnen kann, da sie bestimmte Kriterien erfüllen. Folglich wird die angenommene Hypothese bestätigt und die Bezeichnung ‚künstliche' Intelligenz wird als überflüssig erklärt. Sie ist nur nützlich um die Sonderstellung des Menschen als intelligentestes Wesen des Universums zu halten. Die Autoren der hier behandelten Bücher widerlegen diese Stellung allerdings.