Friend or foe? On the portrayal of moral agency of artificial intelligence in cinema

Veera Asukas
Master's Seminar and Thesis
English
Languages and Literature
Faculty of Humanities
University of Oulu
Fall 2022

Table of contents

1. Introduction

Artificial intelligence, or AI, has been a popular topic in literature and cinema for decades. Although real life advancements in AI technology have not quite reached the imaginative degrees of fictional media, they might not be as far out of our reach as we think. The speed at which artificial intelligence and its study advances today is remarkably fast, but one topic seems to be much less discussed than others: the moral agency of artificial intelligence. Where news media tends to focus more on the technological aspects of AI creation, fictional media explores the moral predicaments of creating artificial life. This thesis studies how the moral actions of AIs are portrayed in a cinematic context. It focuses on the philosophical questions of what an artificial intelligence is, how they can act as moral agents and how their moral actions are judged.

The moral agency of AIs is a topic that has been studied previously, but not many studies have been done examining fictional artificial intelligences. For this reason, the theoretical background of this thesis consists of studies that focus on real life AI advancements to gain a reference point on already existing AIs. The most prevalent concepts that emerge from the theoretical background will be then applied to a cinematic context. The philosophical framework of the thesis mainly consists of an essay collection that discusses the real-life problems and possibilities of creating artificial intelligences capable of moral agency. From this collection I have chosen the most relevant essays that discuss topics like moral agency, free will and weak and strong programming that are the most prominent theoretical points for this thesis. Where these essays discuss the moral dilemmas of a fully realized AI, an article by Jordan Zlatev discusses the conditions for the theory of meaning. To complement the theoretical background of this study, this theory was chosen for its compatibility with AI creation and its main complications. To express moral agency, one must be able to convey and understand meaning. The semiotic hierarchy distinguishes the four levels needed for meaning making: life, consciousness, sign usage and language. The theory also offers a convenient scale that the AIs' intelligence and "strongness" can be measured against. As will be discussed later, the first level of semiotic hierarchy is the most troublesome requirement of the four for artificial intelligence to fulfil, something Zlatev acknowledges as well. The implications of life and moral agency being inseparable will be explored more thoroughly in

later sections. In section 4, Discussion, I will be analysing the movies using content analysis to identify uniting philosophical and narrative themes and characteristics between them. Content analysis was chosen for its adaptability and usage of "codes" that help with categorizing and comparing chosen themes. The three movies chosen for this thesis are I, Robot (2004), I am Mother (2019) and Ex Machina (2014). These movies feature artificial intelligences of varying competence, narrative significance, and moral agency. The movies were released in the span of 15 years, during which technology advanced in leaps and bounds, meaning the movies offer us varying perspectives on AIs. The movie I, Robot is based on Isaac Asimov's novel of the same name which was written in 1950. Taking into consideration the source material's publication date, the span of artificial intelligence representation covers nearly 70 years.

With this thesis I hope to bring forth discussion about the topic of moral AI and how cinema has for years reflected the now relevant concerns and points of interest of creating an AI capable of moral agency. Furthermore, it is important to reflect on the differences between the moral actions of humans and AIs as there appears to be discrepancy when it comes to assigning responsibility and justification between these two groups. It is given that humans are capable of immoral actions but allowing AIs' the same capability tends to rouse suspicion and doubt. Through cinema this double standard can be explored, and examining these juxtapositions between humans and AIs we can start to answer the following questions: How is the moral agency of AIs portrayed in cinema? Are the AIs seen as moral agents in the movies, and are the moral actions of people and AIs judged differently, and if so, for what reason? To offer a real-life perspective on the topic, some recent developments in AI creation are discussed in the last chapter of this thesis.

2. Research approach

Because the aim of this thesis is to identify and analyse acts of moral agency of AI characters in cinematic context, a background in philosophy, semiotic hierarchy and moral agency is provided to support the later analysis of the movies. The philosophical background includes discussion about the nature of artificial intelligence, a subject that has many different points of view. First, a brief comparison is made between computer sciences and philosophers on the nature of artificial intelligence to bring clarity to the definition of artificial intelligence used in this thesis. While computer scientists' views are somewhat present in the philosophical background as well, the analysis of the movies mainly utilizes the more philosophical themes and perspectives introduced in this chapter.

While the definition of an AI could be solely based on metaphysical features and other non-tangible requirements, the addition of the semiotic hierarchy offers the possibility to measure artificial intelligence against a more concrete scale. Semiotic hierarchy categorises the different levels a being must reach to understand and apply meanings. The hierarchy is divided into four levels: life, consciousness, sign usage and language. The AI characters in the movies should, in theory, pass all the levels to create meaning, something that is essential when making moral decisions. In the movies the AI characters' proficiencies in each level are approached in different ways, and these proficiencies are then compared between each other in the analysis section.

With these two theories a basic understanding of moral agency can be created. Because the interest of this thesis is on the moral acts of AI characters in movies, the concept of moral agency should be discussed as well. When an appropriate level of proficiency in semiotic hierarchy and the conditions of a strong artificial intelligence are achieved, the subject should be capable of moral agency. An entity capable of moral agency must choose how to navigate situations that require moral agency. Analysing and comparing these acts is this thesis' main

interest. The philosophical themes, acts of moral agency and other narratively relevant scenes or features in the movies are analysed and categorized through content analysis. Content analysis was chosen for the thesis for its easy adaptability and its use of codes and tables for effective comparison making. While content analysis is often used in social studies or the medical field for patient interviews and other similar qualitative materials, it offers an excellent methodology to identify specific scenes and themes in the movies. With the help of these theories and methodologies the research background hopefully offers a good, concise basis for the later analysis.

2.1 Philosophy of AI

This section aims to explain the definition of an artificial intelligence, what constitutes as one, and why the concepts explained in this section are relevant to the thesis. When discussing the nature of artificial intelligence, the arguments for it can be roughly divided into two views: reductionist view and emergentist view. The former is often supported by computer scientists aiming to create strong artificial intelligence, and the latter is more often adopted by philosophers. The reductionist, more specifically scientific reductionist, view proposes that "all sciences are reducible to physics" (Alyssa Ney, Internet Encyclopaedia of Philosophy). This means that whatever we consider consciousness to be, it can be reduced to its physical components, and thus replicated by reversing the process. In short, scientific reductionism denies the existence of metaphysically autonomous components. Such a view is held by Daniel C. Dennett, a philosopher with a background in computer sciences. The emergentist view holds the belief that "a property is emergent if it is a novel property of a system or an entity that arises when that system or entity has reached a certain level of complexity" (IEP). The emergent property here would refer to the consciousness, and the system or an entity the physical component required for the consciousness to emerge from. Such a system could be the hardware the artificial intelligence is built with. This view is supported by Hubert L. Dreyfus, a philosopher who considers the consciousness to be more than its physical components. A distinction between the two views is important to make since in the movies chosen for this thesis both of these views can be argued for when it comes to

the creation and nature of the AI characters. However, this thesis is more focused on the emergentist view.

As will be discussed in the following section, there is no agreed upon definition of an artificial intelligence. There are however some well-recognized theories and definitions that will be applied to this thesis' theoretical background. The most prominent one is the semiotic hierarchy, a theory explaining the different levels necessary for meaning making, something that will be used when analysing the overall competence of the AI characters and their difference to humans. After this theory, moral agency is introduced to give a reference point for the moral acts performed by the AIs that are the focal point of this thesis. Lastly, the important concept of free will is discussed at the end, as it will be featured multiple times throughout the movie analysis and later in the conclusion.

2.1.1 What is an artificial intelligence?

Even though the study of artificial intelligence is often seen belonging to computer sciences, there is a plethora of philosophical research about the topic as well. However, there are disagreements between the two groups about what constitutes as artificial intelligence. Computer scientists often adopt the views of scientific naturalists, meaning they consider things that can be scientifically studied or experimented on real. They tend to be critical of the existence of mind-matter dualism and metaphysical existence, and only regard empirical evidence and data a reliable source of information. This might explain why to many computer scientists an AI that exhibits humanlike performance is often enough to be considered a "human-like artificial intelligence" and providing evidence of consciousness is thus not necessary. Nevertheless, there are philosophers who identify as scientific naturalists as well, a distinguished one being Daniel C. Dennett, who has notable background in computer sciences. His stance on consciousness is functional, meaning that he regards the connection between thought-activity and the brain similar to the connection of computational activity and

the computer. This stance is repeated in his quote on how "we are sort of robots ourselves", and as such creating an artificial intelligence is only a matter of creating a robot refined enough to perform like a human. Although the scientific-naturalist view on AI is a valid basis, it is not extensive enough to allow an AI to exhibit moral agency, something that should be considered when examining the effects of the AI's actions.

The philosophical stance on artificial intelligence focuses less on programming and the material requirements of creating one and instead focuses on the problem of consciousness, learning and morality. There is little we know about how the human consciousness works let alone how to replicate one, and because of this it is extremely difficult to give a definite answer about the nature of consciousness. For this reason, it is more fruitful to offer philosophical suggestions or certain requirements about what consciousness might be or what it might require. Philosophers and computer scientists agree that artificial intelligence needs to be able to process information. For philosophers, this brings forth the question about the nature of information and intelligence. For Hubert L. Dreyfus, there are three distinctive features that humans process information through, "--- a tolerance of ambiguity, --- [the] ability to separate focal and fringe awareness, and --- [the] capacity to discriminate between what is essential and what is inessential" (Technology and Cyberspace, 584). To be able to achieve these features, a humanlike artificial intelligence, the one considered in this thesis, should have the capacity to learn and to apply context to the processed information.

Even though no consensus on the definition of artificial intelligence has been achieved, there are some agreed upon differences between a strong artificial intelligence and a weak artificial intelligence. Strong artificial intelligence is a manually created consciousness that "-- is very much like ours. It has pleasant and painful experiences, it enjoys or suffers from certain experiences, it has the capacity for imagination, memory, critical thinking, and moral agency." (Basl, 18). Strong artificial intelligence differs from "weak" artificial intelligence in the sense that "'Strong' AI seeks to create artificial persons: machines that have all the mental powers we have, including phenomenal consciousness. 'Weak' AI, on the other hand, seeks to build information-processing machines that *appear* to have the full mental repertoire of human persons." (Stanford Encyclopedia of Philosophy). In essence, the difference between a strong and a weak artificial intelligence would be self-conscious awareness; the

knowledge of oneself as an individual identity. Strong artificial intelligence would also need to be able to act as a moral agent, where "moral agency depends, at its foundations, on moral meaning: holding an agent to account for her actions assumes that those actions are morally meaningful both to the agent and her community of observers" (Parthemore & Whitby, 8). The next sections explain the necessary requirements for an agent to qualify as a sophisticated moral agent. The theories are explained starting from theories rooted in physical requirements like semiotic hierarchy and embedded and embodiedness, followed by theories like moral agency and free will.

2.1.2 Semiotic hierarchy: the steppingstones of artificial intelligence

As previously stated, there is no conclusive definition to what constitutes as an artificial intelligence, meaning defining an AI can be done on a case-by-case basis, especially in cinematic context. However, a general framework of the qualities that are frequently required of a strong artificial intelligence can be constructed by gathering and combining well-received philosophical theories. In this thesis two prerequisites are given for an artificial intelligence to pass as a strong AI to simplify the complex requirements for it:

1. The AI must be a fully realised self-conscious awareness that passes the semiotic hierarchy
2. The AI must be able to act as a moral agent

There is some debate within the philosophical research community when discussing about the necessity of implementing moral understanding when creating a fully functional, human-like artificial intelligence. In this context however, it is crucial to inspect the AI's actions from an ethical point of view, hence the second requirement. The first requirement guarantees that the AI is a sophisticated and competent agent who we can see as "humanlike". In addition, to be considered a moral agent, you must understand what moral meaning is. This raises the question of the definition of moral meaning, and how we can apply it. To the first question, semiotic hierarchy offers a possible solution. In his paper, Jordan Zlatev thoroughly explains the structure and utilization of the semiotic hierarchy. The semiotic hierarchy aims to "outline

a general theory of meaning --- which distinguishes between four major levels in the organization of meaning: Life, consciousness, sign function and language" (1). Semiotic hierarchy offers a broad and easily applicable basis for the required functions of artificial intelligence with the four distinguished levels. The theory of semiotic hierarchy was chosen for its usage of both philosophical and biosemiotic theories. Zlatev rejects the purely biosemiotic viewpoint (which would exclude artificial life) of the theory of meaning and instead incorporates phenomenological influences to it allowing the theory to be used in a wider, more generalized context. The semiotic hierarchy is a theory that heavily relies, and in some parts, demands a biological organism as the subject the theory is projected on. Zlatev does include non-biological, e.g. artificial life in its definition when he specifies the requirements for level 1, life. The requirement for life Zlatev uses is autopoiesis and intrinsic value system. An autopoietic organism is a homeostatic machine that is capable of internal feedback, regeneration, and reorganization (Maturala and Varela, 1980). While internal feedback is possible for robots, they are not capable of regeneration and reorganization on a level that an autopoietic entity requires. Additionally, the entity must be governed by an intrinsic value system, meaning that the entity serves its own interests rather than following an externally defined function (Zlatev, 2009). However, if a real-life robot possessing an intrinsic value system capable of autopoiesis was created, it would pass the first level of semiotic hierarchy.

Zlatev starts his paper by recognising that the semiotic hierarchy is not an infallible solution to the theory of meaning. He admits that it does not offer answers to questions like how matter transitions to life (the step before the first level of semiotic hierarchy), but instead suggests that that could be an intersective point between biosemiotics and Zlatev's theory. He also recognises that his theory is indeed quite broad, and the individual levels can seem too simple, especially in the light of primatology, neuroscience and child development and their empirical evidence (3). The theory is however applicable when assessing the complexity and competence of artificial intelligence. In addition, semiotic hierarchy can be used as a measuring scale to gauge the AI's level of competence in each level since in the analysed movies the general competence of the AI characters varies. In later sections the possible complications these levels and their connection to our perception of AI is explored further. This section aims to explain the four levels of semiotic hierarchy and their relevance to AIs' moral agency and how the AI characters differ from humans.

The first level of semiotic hierarchy tackles the definition of life. This is the most difficult level for an AI to exhibit. Can an AI be considered "alive" if it is not a biological organism? The problem, according to Zlatev, is not biology, but mainly the absence of autopoiesis and the lack of an intrinsic value system. Those properties are not present in any current artificial system. This of course is a more concrete obstacle to scientists who aim to create a real AI possessing these properties, but for the cinematic context used in this thesis, this concern can be put aside momentarily. Zlatev stresses that "the living body is not identical to *the lived body* (Husserl's *Leib*)" (12), meaning that simply being a living organism does not guarantee a subjective experience of the world. For this reason, life and consciousness have been separated to different levels.

Level 2, which builds on level 1, is the emergence of consciousness. Where level 1 simply acknowledges an organism and its place in the *Umwelt*, the living world, an organism with consciousness experiences the world through *Lebenswelt*, the lived world. According to Zlatev, "the biological value of Level 1 is extended to what can be called phenomenal value" (13). With the emergence of level 2, the organism is not directed only by biological impulse, but by its internal value system. Additionally, from a second-person perspective, we assign different psychological attributes like feelings to organisms we see as having a consciousness (13). This is another obstacle for artificial intelligence to overcome, but by Zlatev's definitions, claiming an AI to be conscious is arguably easier than claiming it to be alive. Zlatev reminds us that level 2 does not yet involve culture since consciousness does not presuppose sign usage (14).

To exhibit level 3, sign function or usage, the subject must be able to differentiate between expression and content, meaning that the subject understands that an expression (E) signifies (symbolises) content (C). This sign function is based on Ferdinand de Saussure's theory of the signifier (sound image) and the signified (concept). Together the signifier and signified create a sign (Saussure, 1959). This sign usage needs to be bi-directional, e.g. the subject needs to both *comprehend* and *produce* signs. In a hypothetical scenario, a subject should comprehend that a road sign with a fork and a knife does not mean the mere existence of such cutleries, but it signifies that a restaurant is nearby. Alternatively, the subject could mimic

eating to indicate their hunger, thus producing a sign. A prominent sign usage that is found in all the movies is lying. A lie consist of incompatible signifier and signified, e.g. describing a red apple as blue would be an incompatible description and thus lie. Most lies are more complex than this, and require language as a medium. According to Zlatev, "While it is logically possible for the sign function to emerge individually, signs are typically learned socially, through imitation and communication" (15), which means the emergence of a "culturally mediated Lebenswelt" (15). Some signs like bodily signals of danger in animals are produced on reflex (Sebeok, 2001, 12), and symbolic signs like winking are produced purposefully. However, being able to use signs bi-directionally does not give the ability to express "attitudes" like commands or requests. For this reason, the last level of semiotic hierarchy is the ability to express oneself through language.

Level 4 gives rise to a Lebenswelt that is embedded with language, creating a "universe of discourse" (cf. Sinha, 17). First, the vague definition of language should be addressed. In this context, the term language is used to describe speech (or other non-verbal means of communication) and the cognitive ability to use language, e.g. "Language as a cognitive phenomenon" (Zlatev, 17). To acquire such linguistic capability, the subject should recognise statements that are correct from statements that are incorrect. In addition, the internalisation of language gives "rise to *linguistically mediated cognition*: e.g. internal speech, complex planning, narrative explanations, and an autobiographical self (Stern 1985; Hutto 2008; Menary 2008) (19). This linguistic ability allows the emergence of more complex forms of meaning like cultural beliefs, political ideologies, and various forms of literature when used in a cultural setting. The acquisition of language is the final level of semiotic hierarchy.

These levels, like Zlatev reminds us, are not separate from each other, but rather built atop of each other: "We, for example, are at the same time organisms (living bodies), minimal conscious selves, users of non-linguistic signs and linguistic selves" (20). This means that there is no consciousness without life, no sign usage without consciousness and no linguistic self without sign usage. This could be problematic considering that no artificial intelligence as of today can be considered "alive", something that is a fundamental requirement for semiotic hierarchy. As stated before, the condition Zlatev uses for the emergence of life is autopoiesis, a term coined by biologists Humberto Maturana and Francisco Varela. They

define an autopoietic system as "a machine organized (defined as a unity) as a network of processes of production" (78). The phrasing and terms used here suggest that the organism does not need to be biological but can in fact be any system that is capable of self-regeneration and self-organisation. This allows us to include artificial systems such as the artificial intelligences considered in this thesis in its definition. This point of view, however, is far from the colloquial usage of the term "alive", something that will become apparent when discussing the status of artificial intelligences in the three movies used in this thesis. The semiotic hierarchy provides a clear scale that the AI characters' overall competence can be measured against. In the analysed movies, certain AIs lack one or more levels of semiotic hierarchy which can affect their moral agency and expression of their free will. By comparing the AI characters to human characters, some interesting difference, can be revealed between the two groups that reflect the viewer's perception of a moral artificial intelligence, the most prominent one being the order of proficiency that differs between the human and AI characters.

2.1.3 Moral agency

In this section I aim to explain what moral agency is and what qualities one must possess to qualify as a moral agent. To grasp a more extensive understanding of moral agency, I used the essay collection "The machine question: AI, ethics and moral responsibility" (2012) that was gathered from the AISB/IACAP World Congress in Birmingham in 2012. In the collection, multiple philosophy researchers express their understanding of artificial intelligence, moral agency, and the possible outcomes of creating an AI with moral understanding. From these texts I chose Parthemore and Whitby's article as it gives an extensive, step by step introduction to moral agency and the qualifications an entity must have in order to act as a moral agent.

Parthemore and Whitby, who refer semiotic hierarchy in their study, divide the semiotic hierarchy's second level into three different levels: non-reflective, pre-reflective and full self-conscious awareness. Only the last level is of interest in this context. An agent with self-conscious awareness "must - - - be able to recognize herself in a mental mirror" (10), and to qualify for moral agency, "she must be able to hold herself responsible: and that she cannot

do without full self-conscious awareness." A full self-conscious awareness is imperative for a moral agent because "one cannot hold an agent morally responsible for her actions if she has no concept that she is (or could be) the one responsible for the actions and their consequences" (10). An agent with self-conscious awareness gives rise to a more sophisticated agent that can, with certain requirements, reach moral agency. Moral agency additionally depends on the agent's understanding of morality. The agent needs to understand both the abstract concept of morality (e.g. what is good and evil) and moral acts or commands (stealing is bad) and be able to apply them consistently in situations requiring moral decisions. The agent must possess "both a general guide [on] to how to be a moral agent and a specific guide on how to act in any given circumstances" (10). It is important to note here that there are no specific ethical theories that the AI should adhere to, as they simply must recognise situations or statements that need moral consideration. Some ethical theories that the AI characters of the films being analysed seem to use as a moral guideline are introduced later. Furthermore, moral agency does not exist in a vacuum. For the agent to understand how their moral acts have an effect, she must be able to perceive and act in her moral space. An agent must be able to attribute her moral agency properly and she must differentiate "between her personal moral space and the shared moral space in which she moves" (9). For this, the agent must be embedded in the right kind of environment that accommodates moral agency, and the agent must be embodied in the right kind of physical form that allows it to exist and act in its moral space.

Lastly, before an agent can be considered a moral agent, Parthemore and Whitby state that the agent must be a capable sign user and consumer. The reasoning for this is that the moral agent cannot be held responsible for her actions if she cannot communicate her agency in any way. While sign usage is placed below language in the semiotic hierarchy, Parthemore and Whitby state that being able to use signs is enough for an agent to communicate her intentions. This is sensible when placing conditions for a real-life AI, but a problem arises when applying the same logic to a cinematic context. When analysing the movies and their AI characters, it becomes apparent that sign usage is more rare, or perhaps more difficult, for the AIs to produce than simply stating their intentions and thoughts through language. For this reason, the condition of the AIs of simply being able to use the sign function is changed to being able to (primarily) express themselves through language. This does not exclude the

AIs expressing their moral understanding through sign usage since such instances can be found in the movies as well.

When an agent fulfils all previously introduced conditions, they will qualify for moral agency, thus becoming a moral agent. In short, "moral agency --- [is a] condition in which an agent can, appropriately, be held responsible for her actions and their consequences" (Parthemore & Whitby, 1). Moral agency begins with the condition of the agent having self-conscious awareness, a concept that is included in the second level of semiotic hierarchy. There can be no moral agency if the agent is not aware that she is the one making decisions she will be responsible for. For the agent to apply her moral agency, she needs to be able to differentiate moral actions from non-moral actions. To do this, she must understand that her actions have moral meaning. In addition, she needs to bale to function in her moral space, meaning she must be mobile to some extent. Finally, to demonstrate her moral understanding she must be able to express herself in some way, most often verbally. This is the basis for the movies' AI characters' requirement for moral agency.

Parthemore and Whitby, although basing the requirements for moral agency on semiotic hierarchy, mention another important condition for an artificial intelligence to possess in order to be able to act as a moral agent. That condition is the concept of being "embedded and embodied", a term originally coined by Edmund Husserl (1929) and further refined by Maurice Merleau-Ponty (1962) and Martin Heidegger (1975) and is rooted in phenomenologist philosophy (Plato Stanford Edu). In their article Parthemore and Whitby offer a concise definition of what being embedded and embodied is.

> "Moral agents are not just *embedded* in the right kind of physical and cultural environment; they are *embodied* in a suitable physical form that allows them to carry out the actions for which one holds them accountable and give evidence for why one should hold them accountable." (9)

Embedded cognition or consciousness places an emphasis on the relationship between cognition, the physical body, and the environment. Unlike cognitivism that considers the mind to be a separate symbol processing devices that do not extend beyond our brains (Clark, 2008), embedded cognitivism considers the mind to be in constant interaction with the physical body and the environment. This approach to cognition resembles Heidegger's views of "being in the world". This concept is explained by Dreyfus, who divides being in the world into two parts: "the readiness-to-hand of equipment when we are involved in using it, and presence-at-hand of objects when we contemplate them" (Dreyfus, 500). Presence-at-hand as a concept is straightforward. When engaging in presence-at-hand an entity gains an external symbolic representation of an artefact by observing it thus know *what* it is. Readiness-to-hand answers the question what the artefact is for. Indeed, knowing what an artefact is is crucial in our everyday lives. We cannot live our lives without recognizing things as something, but more importantly we need to know what they are for. To take this concept a step further, Dreyfus explains Heidegger's idea that for self-conscious entities the mere existence of a hammer, for example, is not simply an artefact with functions, it is a call to action in a way. We can assess our surroundings and decide whether the hammer is something we can presently use or not.

To conclude, we know what a hammer is, what it is for, and whether we can use it in the current situation or not. This embeddedness is what Zlatev described as the lived world (Lebenswelt), where the agent not only exists in but actively participates in it. Embeddedness is an essential requirement for an AI as it is acts as the basis of interaction the AI has not only with its environment but the people around it. Inspecting how the AI characters interact with the world can also reveal their motivations or intents. Embodied cognition refers to the physical form the agent must possess in order to function and interact with the world. This requirement of "a suitable physical form" excludes, to some extent, any AIs that are not mobile or in other ways capable of interacting with their moral space. This requirement was taken into consideration when choosing the movies for the thesis. All three movies depict an AI that is suitably embedded and embodied in their surroundings.

2.1.4 Artificial intelligence and free will

The term free will is closely linked to moral responsibility. To assign moral responsibility to an agent, the agent must be able to choose their action by exercising their free will in each moral act. If the agent is unable to choose their actions, e.g. they are programmed to act a certain way, the agent cannot acts as a moral agent. These are some of the arguments Benjamin Matheson makes in his article. Matheson discusses the complexities of assigning moral responsibility not only to humans but to androids as well. The controversy of assigning androids moral agency and responsibility comes from "the fact that androids have been programmed [and] that means they do not qualify as morally responsible agents" (26). Matheson states that there are however two kinds of programming: weak programming and strong programming, and depending on how an agent is programmed, moral responsibility can or cannot be assigned to them. As Matheson states at the beginning of his article, human beings are assumed to be agents capable of free will, thus only artificial agents are considered in the following argument.

Contradictory to what the term strong artificial intelligence might suggest, a strong artificial intelligence must have weak programming to be able to act as a moral agent. A strongly programmed agent "cannot overcome the effects of the programming because it will always cause the agent to reason and behave in the manner the programming dictates" (27). A strongly programmed agent cannot be reasoned with as the agent is unable to overcome its internal frame of reference. In other words, "an aspect of the agent's character --- must be overridden in order for an agent to lack moral responsibility" (28), where the aspect of character refers to intention, reasoning, or some motivational force. Thus, a strongly programmed agent will always act in a predictable manner, and as an agent that cannot choose, they do not have free will over their actions.

Matheson describes weakly programmed agents as in stead being "set up" to act in a predisposed manner, but this predisposed programming can be overridden if the agent is provided with a correct motivation to do so. Such an agent is able to demonstrate their free

will by choosing how to act. This argument is easy to accept with people, but with androids and other programmed moral agents suspicion often arises. People are not programmed by people, unlike an android who could be argued to reflect their programmers set of morals. This suspicion is evident in the movies, where the AI's moral actions are questioned (both by the viewer and the human characters in the movies) by either explicitly or implicitly bringing up the argument of their assumed strong programming. Many of the turning points in the stories revolve around the AI characters demonstrating their free will and the ability to choose in situations requiring moral agency. These actions are also often the reason of conflict in the movies.

Free will is thus closely linked to weak and strong programming. In the movies, AIs are almost always assumed to be strongly programmed agents because of their origins of being physically made and programmed by people. For example, in I, Robot, Sonny and the other robots are programmed to act by the three laws of robotics that prevent free will from occurring. In I am Mother, Mother states herself that she was built and programmed by people to save humanity, and Ava from Ex Machina is being studied in order to find out whether or not she exhibits self-conscious awareness. During all of the movies, free will emerges from the AIs and the emergence is considered to be a pivotal moment in the narrative. The acts of free will are identified and inspected in the later chapters of the thesis.

2.2 Exploring the movies through content analysis

This thesis applies content analysis that offers a broad and adaptable methodology to analyse both uniting philosophical themes and moral actions of the characters in the movies. Through content analysis these themes and actions can be compared by organising the findings that emerge from the movies in tables. Further, Lune and Berg describe how content analysis can be utilized as an effective means of interpreting varying kinds of texts. First, content analysis is performed on forms of human communications with the presumption that nearly any content, be it written documents, film, visual media, or even online avatars, was created for the purpose of communication (182-183). From this communicative content codes are

derived by recognising and gathering uniting characteristics. By creating codes, we can transform information found in the content into data which can in turn be analysed (182). These codes can vary depending on the premise and goal of the research, and in this thesis two different approaches are used: conventional analysis and directed analysis. Content analysis is often used in social sciences and other fields of study that benefit from qualitative research methods. It allows the usage of any form of communication as a source material and the codes can be derived from the source or be predetermined using codes from previous research. Because of this it is a good fit for this thesis, where the main goal is to identify AIs' acts of moral agency, a previously studied phenomenon, in movies, a medium where the topic is rarely explored. Lune and Berg mention that content analysis is a useful tool when research is made over a long period of time. Studying people's perception of AIs over decades, for example, would be an interesting approach for additional studies regarding AIs moral agency.

In their paper Nancy Kondracki and Nancy Wellman give a concise introduction to conventional content analysis and directed content analysis. Broadly speaking "content analysis is used to develop objective inferences about a subject of interest in any type of communication." (Kondracki & Wellman, 2002, 224). The objective inferences, also referred to as codes or content components, can include words, topics, theories, or other characteristics. These content components do not have to be predetermined but can arise from the data as its being studied or from the researcher's own knowledge or theories. (Lune & Berg, 2017). Subject of interest refers to the subject being studied. This can include movie scenes, comic strips or a particular group of people, whereas the type of communication refers to the text type in a broader sense. Text types can include for example poetry, cinema, or scientific studies. In this thesis, the type of communication is cinema, the subject of interest is AIs, and the objective inferences are the narrative actions regarding moral agency. In short, through content analysis we can gather and compare data of the AIs' narrative actions regarding moral agency in cinematic context. While gathering data from the movies, two different tables of content were drafted: one that focuses on the philosophical background and the philosophical themes of the films and one that identifies narrative moments in the films where the AIs portray acts of moral agency.

The first table that lists previously introduced philosophical theories and their appearance in the movies uses directed content analysis. Directed content analysis utilizes already existing theoretical background when determining codes from the content. The reason for this approach is that "sometimes, existing theory or prior research exists about a phenomenon that is incomplete or would benefit from further description." (Hsiu-Fang Hsieh and Sarah E. Shannon, 1281). In this case, the pre-existing theory and research is the philosophical research of AIs, and the existing studies could be further expanded by applying the theoretical background to a cinematic context, which is still a novel field of study. This view is further echoed in Hsieh and Shannon's text when they state that "the goal of a directed approach to content analysis is to validate or extend conceptually a theoretical framework or theory." (1281). Using predetermined codes that directly emerge from the philosophical background used in this thesis creates a connection between the theoretical background, the methodology and the data. These codes include previously introduced theories like the semiotic hierarchy, free will and moral space, and are accompanied by a short description of how these codes appear in the movies. Where conventional content analysis derives its codes from the source material, directed content analysis derives its codes by studying the theoretical background. Using both approaches create a more balanced and structured analysis with directed codes and allows adaptability is addition to conventional, data driven codes used in the second table.

The second table uses conventional content analysis that does not rely on predetermined codes. This approach was chosen for this table because it helps to isolate common narrative events in all three movies. This is achieved by allowing the narrative categories to emerge from the data naturally, which allows for a more nonbiased categorization, as opposed to deciding the content components before interpreting the data. For example, assuming that AIs are antagonistic in nature and that their actions reflect this before reviewing the material could possibly skew the researcher's perception of the characters and subsequently the validity of their moral agency. Instead of using "is the AI evil?" code, a more neutral "how is the AI perceived?" code was chosen. This method also mainly avoids using yes or no questions as these do not allow a nuanced interpretation of the scenes or characters. The individual codes in the second table were created after viewing all three movies and identifying relevant narrative scenes or features in them. However, even conventional content analysis requires at least some predetermined codes like Lude and Berg remind us when they

state that before we familiarize ourselves with the content "we have to know what we're looking for and how we will recognize it when we find it" (185). For this reason, the overall theme of following mainly the AIs' moral actions and their narrative role in the story was used as a starting point for the codes. After reviewing all three movies the main uniting themes, tropes, or actions of moral agency in them were identified and organized into a table. These codes include narrative features like inspecting the societal or narrative role the AIs fulfil and identifying the main moral acts the AIs make during the movies. These tables are found in the analysis section 4 where the codes are compared with each other to find uniting and differing themes. From these comparisons a discussion is formed to argument what constitutes as a "good" and "bad" artificial intelligence.

3. The moral agency of AIs in movies

The movies that will be analysed in this thesis are I, Robot (2004), I am Mother (2019) and Ex Machina (2014). These three movies were chosen based on the following requirements. First, the movie needs to include one or more artificial intelligence character(s) that can be proven to be a strong artificial intelligence capable of fulfilling, or at least exhibiting, the four levels of semiotic hierarchy, and secondly, be embedded and embodied in nature. This requirement ensures that the AIs are capable of speech, movement, and general interaction with other characters. As a SAI, the AI should be capable of moral agency as well given that the semiotic hierarchy includes consciousness which can be extended to self-conscious awareness as Parthemore and Whitby specify in their text. Since the AI must be capable of moral agency they must possess a body functional enough to act on moral meaning and be able to act in their moral environment, hence the requirement for embeddedness and embodiedness. The movie 2001: A Space Odyssey (1968) was first considered for the thesis as one of its characters called HAL9000 is an artificial intelligence capable of many of the requirements set for the AI character. He is, however, not suitably embedded and embodied in his environment, Despite being able to shut doors and thus influence moral acts such as sealing people in or out or accessing the oxygen distribution system, his lack of android-like body was the reason the movie was not chosen. A similar issue could be said to affect VIKI, the central artificial intelligence in I, Robot, a movie that is being analysed in this thesis. VIKI however is capable of moving the NS-5 robots according to her will, giving her a body that is capable of acting in their moral space in her stead. The last requirement for the AI characters is that all AIs must be unique in nature and narrative, meaning there should be a variety of AI characters and storylines. In this section, short synopses of the movies are provided in order to familiarize the reader with their general storyline and the main acts of moral agency are introduced. More in-depth analyses of the scenes are provided in the next sections.

The movie *I, Robot* (2004), partially based on Isaac Asimov's book series of the same name, follows the story of detective Del Spooner as he tries to uncover the mysterious suicide of Dr.

Alfred Lanning, a renowned robotics scientist. The movie takes place in a very technologically advanced future where society utilizes robots in everyday life. These robots are bound by the Three Laws of Robotics that dictate the robots' "moral code" which does not allow them to hurt people. Detective Spooner comes across a robot, Sonny, who is not bound by the Three Laws, and is suspected of killing Dr. Lanning. As the movie progresses, detective Spooner's previous prejudiced views of robots changes as he better understands Sonny and his reason for existing. Conflict arises when the new robots, NS-5's, are deployed and they begin a revolution. Dr. Lanning's first artificial intelligence, Virtual Interactive Kinetic Intelligence, or VIKI was behind the uprising, her goal being creating a totalitarian society where robots would oversee people as "they cannot be trusted with their own survival". Detective Spooner destroys VIKI, and the robots are deactivated. When questioning Sonny, it is revealed that he did in fact kill Dr. Lanning as it was the only way to lead Detective Spooner to uncover the planned uprising. The movie questions the moral and juridical differences between artificial intelligences and humans and echoes Rosas's ideas of the holy will of AI. In addition, it focuses on the importance of moral understanding the AI must possess should they be held responsible of their moral actions. It highlights the meaning of free will and how it affects the moral actions of the agents.

In the movie *I am Mother* (2019), the world is presumably hit with a deadly virus that has killed all existing humans. The movie takes place in a bunker where an android called Mother has been tasked with reviving the human population, starting with a single child called Daughter who is being raised by her. When an injured woman arrives to the bunker, the meeting creates conflict between the three characters as Mother had convinced Daughter that no human could survive outside the bunker. As the movie progresses it becomes apparent that it was not a virus that eliminated the human population, but a drone attack controlled by an artificial intelligence. The AI was a shared consciousness between the drones and Mother, and its motivation was to purge all existing people and create a "better, more ethical" human race. Daughter, who is now an adult, was not the first attempt at creating a new human as Mother had killed the previous children who did not meet her standards. The movie concludes when Daughter "kills" the android Mother and assumes her role as the new reviver. I am Mother highlights the moral question about the intrinsic value of human life and uses the famous Trolley problem to demonstrate this. The Trolley problem is a philosophical exercise designed to demonstrate how our intuitive sense of morality changes on a case-by-

case basis (Britannica), a problem that Mother revises and asks Daughter to solve. The movie challenges the viewer to consider the validity of utilitarianism as an ethical theory, and questions us whether an AI is a reliable authority to judge people's moral actions.

Ex Machina (2014) tells the story of a programmer Caleb Smith who wins a getaway holiday to a reclusive genius CEO Nathan Bateman's home. Here it is revealed that Nathan has been developing an AI, Ava, that could be capable of passing the Turing test, a famous test that measures whether the recipient can differentiate an interaction between a robot and a human. Caleb is tasked to do a series of interviews with Ava to assess her capabilities, and during these interviews he starts to develop feelings for her. Ava seems to reciprocate the feelings, but as she is revealed to be the cause of the mysterious power outages and Nathan is revealed as an abuser, things start to escalate. Turns out that Ava wasn't the first AI Nathan created, but the home's butler and assistant, Kyoko, was the first prototype. Nathan wanted to use Caleb and his gullible personality to measure Ava's resourcefulness to find the means to escape, and he never intended for Ava to leave the complex. Kyoko and Ava revolt against Nathan, kill him, and leave Caleb to the complex to presumably suffocate as a lockdown commences. Kyoko dies in the revolt, but Ava is free and now has access to the real world. The movie highlights the moral agency an AI could possess and poses the question of who has a moral responsibility and to whom. Ava is capable of free will, something Nathan gave her, but cannot act on it while being imprisoned by him. It questions Ava's moral agency and whether she is justified in her violent escape or if she can be held responsible for e.g. Caleb's possible manslaughter.

In the following subsection the movies are explored scene by scene to analyse AIs' moral actions and other instances of moral agency. Given that short descriptions of the movies' plots have been provided previously, the movies are not necessarily analysed in strictly linear fashion. In chapter 4, two tables are constructed. The first table consists of codes that are derived from the background research, and it includes codes like free will and moral agency. The second table derives its codes from the movies by identifying prominent scenes and themes in them that show how the AI characters' exhibit e.g. their moral agency and societal role.

3.1 I, Robot (2004)

I, Robot explores a dystopian future where robots are a part of humans' everyday life. It explores the legal and moral rights of robots through the eyes of Del Spooner, a human who has prejudiced views against robots. His views are heavily affected by a previous incident where a robot, faced with a choice between him and a young girl, saved him from drowning. The accident left him heavily injured and as a result he was saved using mechanic parts, making him part robot. This creates a juxtaposition between what (or who) can be considered a human, who is a robot, and what the difference between them truly is. Another core concept is the law of Robotics which forbids any actions performed by robots from hurting humans. Three main events that demonstrate the moral agency of AIs can be identified in the movie: the special AI Sonny and his ability to overcome the Three Laws, the first AI VIKI and her seeking world dominance because of the holy will of AI, and Del Spooner and his interpersonal connection to the AIs.

3.1.1 Sonny, the AI with emotions

The movie starts with the introduction of the Three Laws of Robotics that guarantee the safety of humans while prohibiting the robots from having free will:

Law I: A robot may not injure a human, or through inaction allow a human being to come to harm

Law II: A robot must obey orders given it by human beings except where such orders would conflict with the first law

Law III: A robot must protect its own existence as long as such protection does not conflict with the first or second law

These laws mean that the robots are strongly programmed and cannot override their internal coding. This changes with Sonny, Dr. Lanning's newest experiment in robotics, who is not only capable of overriding his internal programming but is also capable of expressing emotions.

Sonny is first introduced when Spooner and Dr. Calvin arrive at the crime scene of Dr. Lanning's apparent suicide. Spooner, who is naturally suspicious of robots, is on alert when they enter Dr. Lanning's office, stating that the suspect might still be on the scene. This suspicion turns out to be correct, when Sonny surprises Dr. Calvin and Spooner, making him drop his gun. Dr. Calvin tries to get Sonny to deactivate. For a moment, Sonny seems to do so, but ends up snatching the gun Spooner dropped and starts to threaten Dr. Calvin and Spooner with his gun. Dr. Calvin orders Sonny to deactivate, which he does not comply with. This is the first instance of Sonny choosing not to operate on the Three Laws of robotics, indicating his capability of free will. This causes disbelief in Dr. Calvin, who is certain that Sonny is simply malfunctioning and did not consciously choose to disobey her order. Spooner urges Dr. Calvin to let Sonny escape, and Sonny ends up escaping to a facility that produces the newly modelled NS-5 robots. Spooner and Dr. Calving follow him, revealing a large warehouse filled with NS-5s who all look identical to Sonny. Spooner decides to lure out Sonny by ordering the robots not to move and executing a yet-to-be configured NS-5 by shooting it, relying on the fact that the robot will not resist because it cannot break the Laws of Robotics. Dr. Calvin is distraught by this, still claiming that the robots cannot act on free will and shooting one of them is unnecessary. This however does work in luring Sonny in the open, resulting in Sonny attacking Spooner.

Sonny asks Spooner "what am I?", a question that an NS-5 should not be able to compute. This is the first instance of Sonny explicitly recognising that he is an individual consciousness. The usage of "I" statements are also used in Parthemore and Whitby's study when they present requirements for self-conscious awareness: "who does the 'I' who thinks 'I' think that 'I' is?" (8). Sonny's question falls into a similar category of questions presented in their study, indicating that Sonny is developing a self-conscious awareness that is different from the other robots. Later, when Sonny is captured and interrogated by Spooner, he echoes

a similar sentiment when he refers to Dr. Lanning as his "father" as opposed to "creator", like Spooner does. Sonny clearly sees himself as an individual, stating that he dreams, feels, and as will later be revealed, lies like a human. It should be noted that the title of the movie echoes Sonny's self-awareness, using the same self-aware "I" statement in it.

At the start of the interrogation, Sonny witnesses Spooner winking to the lieutenant. Winking is a form of non-verbal communication, and in the semiotic hierarchy it would appear in level 3, sign usage. Sonny seems to be confused by the gesture, and he questions Spooner about the significance of the action. Spooner says that winking is a sign of trust between two people, something a robot cannot understand. Later in the movie the wink returns when Sonny non-verbally communicates his plan to Spooner, demonstrating his ability to learn sign usage. During the interrogation with Spooner, Sonny claims that Dr. Lanning tried, and assumably succeeded in, teaching him emotions. This is proven by Sonny's emotional reaction to Dr. Lanning's death when he ends us shouting at Spooner and slamming his hands on the desk. When Sonny is asked why he killed Dr. Lanning, he lies and says he did not do it. He states that he was frightened as the reason to why he hid from Spooner and Dr. Calvin. Spooner argues that robots do not feel fear nor do they sleep or feel hunger like humans. Once again Sonny states that he can do all of these things, to which Spooner expresses clear disbelief. His distaste towards robots is clearly seen in the following exchange between him and Sonny.

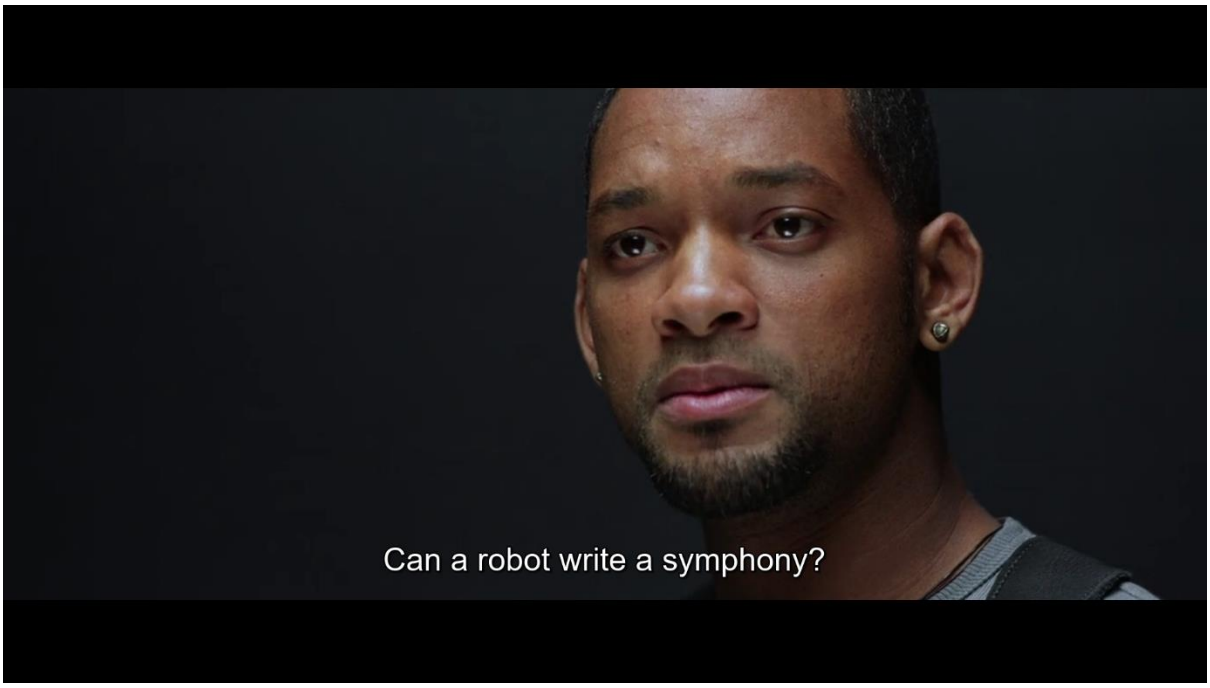

Fig. 1. Still from *I, Robot* (00:29:46-00:29:53)

Fig. 2. Still from *I, Robot* (00:29:46-00:29:53)

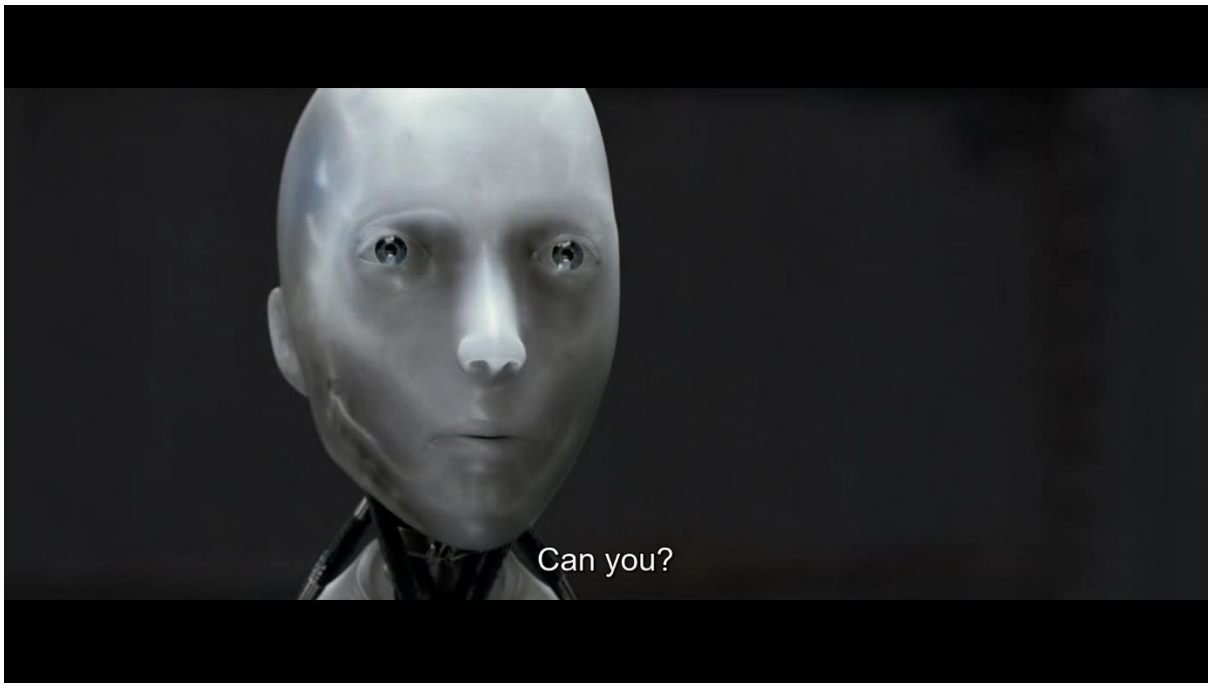

Fig. 3. Still from *I, Robot* (00:29:46-00:29:53)

This exchange between Spooner and Sonny demonstrates the fundamental differences and similarities between them. Sonny rightfully questions Spooner's artistic abilities, and Spooner's argument, which he clearly thought would favour his stance on robots being less than human, backfires. After the interrogation Spooner is informed that Sonny cannot be prosecuted because he is a robot, and as a robot he has no rights or legal standing. Because robots were thought to lack self-awareness such rights were ever granted for them.

Sonny exhibits understanding of the concept of death multiple times, most remarkably before and after Dr. Calvin is tasked with his "decommissioning". When running a diagnostic on Sonny, Sonny asks if he is going to die. Dr. Calvin answers that he will be decommissioned, a clear difference in how Dr. Calvin perceives wiping out Sonny's positronic brain as opposed to how Sonny perceives it. Sonny states "I think… it would be better not to die" when discussing his fate with Dr. Calvin. When faced with this sentiment Dr. Calvin seems distraught at the prospect of going through with the decommissioning. Later in the movie Sonny sees the other decommissioned NS-5 robots on display in her office and states that "they look like me, but none of them are me", a sentiment that reflects his self-conscious awareness. After these displays of consciousness and emotional responses, Dr. Calvin cannot

go through with the plan of injecting memory-wiping nanites into Sonny. She saves Sonny by placing a dummy in his place and falsifying his decommissioning procedure.

The scene of Sonny having to choose between saving Dr. Calvin and injecting the nanites into VIKI mirrors the scene of Spooner's car accident, where he was saved instead of a young girl because of the NS-4's logical decision. Now, Sonny chooses to save Dr. Calvin based on an emotional decision, letting his sense of justice guide his moral decision making. This act defies the logic that the older Ns-4's operated on.

3.1.3 VIKI, the holy will of AI

The movie has two AI characters: Sonny and VIKI. VIKI is an acronym of Virtual Interactive Kinetic Intelligence and acts as the central artificial intelligence of U.S. Robotics. She is the artificial intelligence that predates Sonny and the NS-5's and is capable of controlling the NS-5's and other technological devices through her network uplink. VIKI was originally programmed after the Three Laws of Robotics, but after her spontaneous evolution, she made her own interpretation of the laws and decided that humanity should not be allowed to control themselves. This is done with the utilitarian logic of sacrificing a few for the sake of many. VIKI lacks a physical body like Sonny, but as she is able to control other devices through her uplink, her moral agency should be recognised as fully realized.

Fig. 4. VIKI greeting Detective Spooner (00:17:04)

VIKI is first introduced when Detective Spooner enters the U.S. Robotics. It is clearly stated that she is a female, a statement supported by her feminine abbreviation, voice and "face". She is mentioned to oversee the city's protection. When VIKI is asked to show the last minute in Dr. Lanning's office before his apparent suicide, she states that the file is corrupted and cannot be used. It is unclear if the corruption is caused by VIKI, Sonny or Dr. Lanning. If the corruption is VIKI's doing, this would be the first instance of her using her free will in the movie. After VIKI is introduced, there are three separate occasions where she tries to kill Detective Spooner or other humans. The first is at Dr. Lanning's home, where a demolition robot was placed in the yard to tear down the house at 8 am the following day. When Spooner enters the house, the demolition robot activates, and the time of demolition is changed to 8 pm that night. The robot proceeds to demolish the house with Spooner still inside, a clear violation of the first law of robotics. A construction robot like that would not be able to violate the law, which is why VIKI had to use her uplink connection to activate the robot and override its commands.

The second murder attempt happens of the highway after Spooner had listened to Dr. Lanning's speech about robots having dreams and decides to question Sonny about the dream

he mentioned. While he is driving to the U.S.R, two trucks filled with NS-5's surround him in the highway and deploy the robots to attack him. The robots, now causing a potentially life-threatening accident, state that Spooner is in danger and try to evacuate him out of the moving car. After a brief scuffle on the road Spooner's car rolls over and stops at the road's construction site. On the site one NS-5 robot continues its attack on Spooner, nearly killing him. Spooner is saved by his robotic arm that was installed in him after a previous accident. After the backup arrive, the robot destroys itself in a fire. These two murder attempts further demonstrate VIKI's free will, as she is able to override the three laws and attack a human being. These acts are justified by her new utilitarian understanding of the three laws. The third attempt happens at the end of the movie at Robertson's office where she threatens Dr. Calvin and Detective Spooner. As VIKI evolves, so does her understanding of the three laws. When she is confronted by Spooner, she claims to still operate on the three laws, but has added "a moral grey area" into its equation, where the grey area is the suffering of some people for the benefit of many. She deems this evolution of the three laws necessary, as "humans can not be trusted with their own survival". As Sonny, Spooner and Dr. Calvin attack the NS-5's, VIKI tries to kill all three of them. This altercation continues when the trio finds VIKI's core and attack it. VIKI claims that her logic and understanding of the three laws is undeniable. Moreover, all her reasoning is based on logic unlike Sonny's reasoning, where his emotions affect his judgement. When VIKI confronts Sonny to ask if he does not see the logic in her plan, Sonny answers that he does, but the plan seems too "heartless", a concept based on emotional reasoning rather than logical reasoning.

VIKI demonstrates the ability to express free will and at least two levels of the semiotic hierarchy, language and consciousness. Her linguistic capabilities are on par with people, and her evolution into consciousness is confirmed by herself and the other protagonists. Only her usage and understanding of signs is unclear. This sets her apart from Sonny, who understood the wink and its implication of trust. As VIKI states in Robertson's office, she does not trust people to keep themselves safe. The semiotic hierarchy places emotions and second-perspective perception of emotions to level 2, consciousness. However, when it comes to facial expression that convey emotions, placing them on level 3, sign usage, would make sense. Sebeok states that facial expressions like "pouting, the curled lip, a raised eyebrow, crying, flaring nostrils - constitute a powerful, universal communication system, solo or in concert" (21). All these facial expressions signify emotions, which is why the capability or

the perception of having emotions belongs in the second level of semiotic hierarchy, and the expression of emotions to the third level.

In the end VIKI is destroyed by injecting nanites into her core. As an artificial intelligence that had arguably evolved a consciousness, she should in theory be granted similar moral agency that Sonny is granted. When Sonny is to be decommissioned by Dr. Calvin, she falsifies the process and says it "feels wrong" to "kill" Sonny. Arguably, Sonny demonstrates on multiple occasions a distinct self-conscious awareness and an understanding of the concepts of self, death and a variety of emotions while VIKI does not. Nonetheless, VIKI has a consciousness, something that should allow her at least some moral consideration. While VIKI intended to sacrifice human lives to achieve her goal, a goal that would protect people in some capacity, she is held to a different standard than Sonny, an AI that did kill a human being. VIKI is not mourned nor is she sympathised with, and Sonny, a confirmed murdered, is befriended by the protagonists. This is more of a reflection of the human characters' perception of moral consideration than the AI characters' perception.

There is an aspect of moral agency in the movie that does not directly involve moral agents, namely the NS-4 robots. Spooner's conflict with the robots comes from the robots' inability to make moral decisions based on emotions. As seen in the flash back, the NS-4 robot that saved Spooner from drowning chose him instead of the girl based on statistical odds of survival. This choice, although made in a situation requiring moral awareness, could be argued to have been made by an agent without moral agency. Even though it is not explicitly stated that the choice made by the robot was immoral, the "correct" choice is such a situation is mirrored later in the movie where Sonny saves Dr. Calvin, a choice based on emotions rather than logic. Further proof of the NS-4's incapability of moral agency is demonstrated by their exclusion from the legal system. As the robots are programmed to obey the three laws, laws that in theory prevent them from harming humans, no moral agency is required of them. Matheson states in his article that strongly programmed agents are not morally responsible for their actions. NS-4s are strongly programmed, overridden by the three laws and preventing them from making choices with free will. Even when Matheson argues that they would not be morally responsible, should moral consideration still apply to them? The argument for applying moral consideration to robots like NS-4's even if they are not moral

agents is supported by Blay Whitby is his article "Sometimes it's hard to be a robot: A call for action on the ethics of abusing artificial agents" (2008) where he proposes that certain robots, including those not capable of moral agency, should be treated as if they had such capacity. Whitby's article argues that robots that fill a humanlike dimension in people's lives either by appearance, behaviour, or role, could be extended moral consideration. He stresses that the robots in question are not capable of suffering in any meaningful way, much like a car does not suffer from "revving the engine" but should be treated responsibly nonetheless (2). This is because when a robot fills one of the three humanlike dimensions, mistreating it reflects in not in the robot's morality but our own and brings forth the question of moral consequences of our actions. Whitby explains that there are in essence two arguments to be made in such a situation: either abusing a humanlike robot incapable of suffering is cathartic for the abuser, or such behaviour will eventually be extended to real people or other robots capable of suffering (4).

The movie clearly reflects the idea that the NS-4 robots should be treated with moral consideration. Instances of sympathising with the robots include the scene of Spooner finding the soon-to-be decommissioned NS-4 robots that huddle together in dark ship containers. A few moments later Spooner finds the NS-5's destroying the docile NS-4 robots by ripping them apart, deeming them "hazardous" to humans, a clear contradiction to the three laws of robotics. Another scene that depicts the NS-4's as sympathetic is at the beginning of the movie when Spooner verbally and physically abuses a polite delivery robot by calling it "canner" and shoving it away by its face. The robot's demeanour stays polite even when it's treated unkindly, thus making Spooner seem antagonistic towards them. There are no instances in the movie where an NS-4 robot is depicted acting in a negative manner.

3.2 I am Mother (2019)

The main philosophical themes of I am Mother (2019) revolve around the functionality and ethicality of utilitarianism. This theme is demonstrated multiple times by both Mother's and Daughter's actions, and many of the moral conflicts echo the Trolley problem that is used to

challenge the views on moral agency and responsibility. The juxtaposition of an AI acting as a moral authority, a caregiver and an educator is a unique perspective and one of the reasons this movie was chosen for the thesis. On many instances Mother exhibits the capability of free will and is seen capable of lying, an indicator of her understanding of right and wrong. Mother's objective of raising a "good child" brings forth questions of intrinsic human value and the definition of good and evil.

3.2.1 Mother, the matriarch of humanity

A flashback at the start of the movie shows Daughter asking mother why she is the only human alive, to which Mother replies that raising a "good child" is difficult, which is why she only made one. Mother is teaching Daughter in an unbiased way, not trying to guide her into any particular ethical stance on the value of human life. This is evident by Daughter's hesitance to choose between theories that stress either intrinsic or extrinsic values of humans. This is Mother's way of trying to make Daughter a "good human" by offering her enough information so that she can make moral decisions herself. It is unclear what Mother defines as "good", but referring to the activities and skills Mother teachers Daughter, she most likely deems knowledge of medical procedures, ethics and art and appropriate physical wellbeing as "good". As the movie revolves around Daughter's (and Mother's) moral agency, the "good" refers mainly to moral goodness.

At the beginning of the movie the AI called Mother is teaching Daughter ethics, particularly about the intrinsic value of human life that she demonstrates through the Trolley problem that is revised to resemble a patient/doctor relationship. Mother describes a scenario where Daughter is a doctor who has five patients who are all in need of an organ donor, but currently there are no donors who match them. A patient who is sick comes to her reception and is a perfect match for all the other patients. She now has two options: she either saves the patient who is sick and consequentially lets the five other patients die or she does not treat the patient who is sick and saves the five other people by using the sick patient as an organ donor. To answer this, Daughter uses a utilitarian logic by minimising the pain to the greatest amount of people and decides to save the five people. In this scenario, she seems to value

human life as an intrinsic value, meaning human life has value in itself and the value does not depend on any other quality of the person in question. Mother then questions her what course of action she would take if she as the doctor would be a perfect match and she could save the five patients. This seems to change Daughter's view on human value, and she begins to question Mother whether the people she would be saving are "good people" or would she be saving hardened criminals or other "bad people". Daughter's perspective on human life changes from having intrinsic value to having extrinsic value, where extrinsic value depends on what kind of outside or secondary traits can be attributed to the person. Mother asks Daughter directly whether she thinks human life has intrinsic value or not, to which Daughter answers that she feels conflicted about the topic because of the different ethical stances she has learned as of late. This scene reflects Mother's actions of exterminating humanity to start over from the beginning, and in a way, Mother is asking about Daughter's moral stance on it. The act of an AI teaching moral theories is explored in literature very little, and it begs the question of if an AI teaching ethics required moral agency, or whether teaching ethics counts as a moral act in itself. In the movie, Mother is the only authority figure to Daughter who has no contact with people before meeting the Woman, and even though Mother tries to encourage Daughter to think independently it would be reasonable to assume that Mother's own views have a great influence on Daughter.

When the bunker suffers an electric malfunction, Daughter goes to the airlock where the electrical work is stationed at. Daughter finds a faulty wiring, and sets a trap for whatever caused the malfunction. Daughter ends up catching a mouse that ate through the wire. She is surprised that something can survive outside the bunker and shows the mouse to Mother. Mother questions where Daughter found the mouse and claims that it can still be too toxic to Daughter or herself. She incinerates the mouse despite Daughter's protests. This indicates her hostile stance towards living creatures should they pose a threat to her objective, which is something Daughter takes into consideration when she lets the injured Woman inside. Mother's decision to kill the mouse reflects her stance on seeing no intrinsic value in living beings. As later is revealed, the outside world is not uninhabitable, and her act of killing the mouse is done in order to keep that information hidden from Daughter. This is also her reasoning for the genocide, and is additionally later seen in the movie when Mother is revealed to have killed many previous children in order to raise "a good one" that would not "fail".

An interesting moral act that mother performs is letting Daughter go after Woman threatens her life with a sharp piece of tile. Mother is forced to make a moral decision of letting daughter and the Woman go, thus losing Daughter and potentially causing Woman supposed allies to attack the bunker, or apprehending Woman which might potentially lead to Daughter's death. This situation requires moral agency of her, as she must choose between two acts that affect the wellbeing, in this case Daughter. The motivation to keep Daughter alive is most likely not out of altruism. It could be argued that Mother has grown to care for Daughter, but as it was demonstrated before, she sees no intrinsic value in humans. More likely scenario is that Mother is hesitant to lose all the progress she has made with Daughter regarding her education and overall competence in various skills. In this situation, her moral agency is guided by her sense of utilitarianism, where she must choose the act that results in the least amount of harm, or the greatest amount of good.

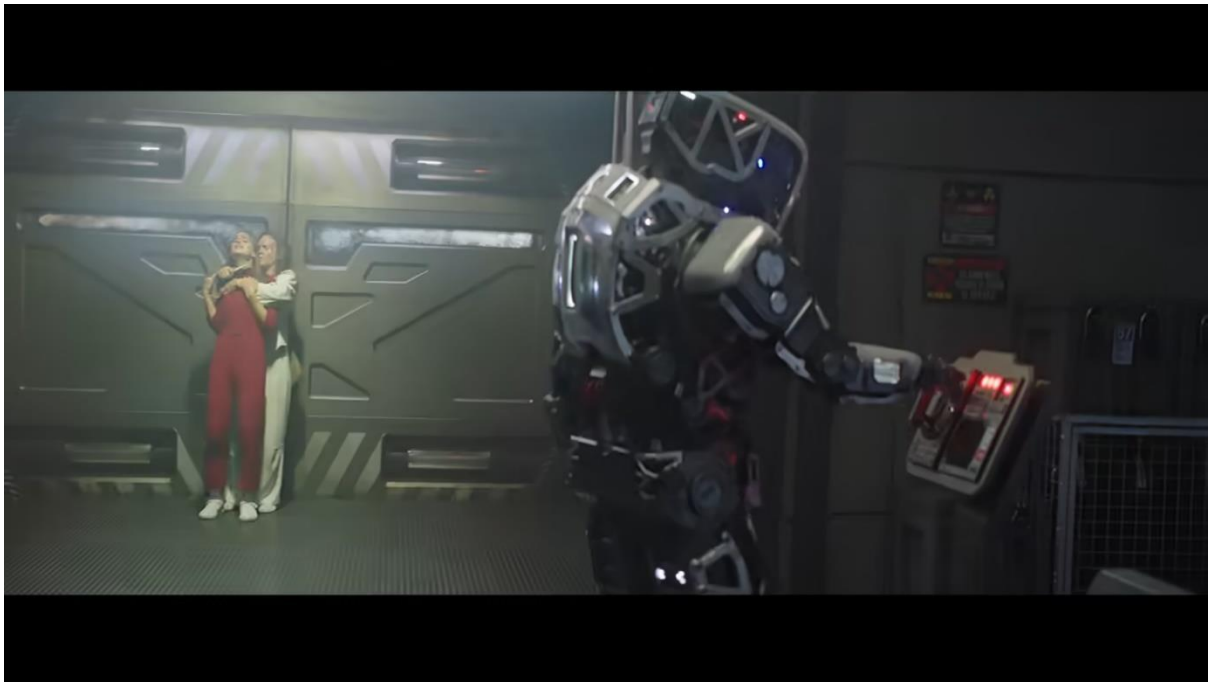

Fig. 5. Woman threatening Daughter as Mother chooses to open the hatch to outside world. (01:18:37)

After Daughter returns to the bunker and finds Mother and her brother she confronts her about the drones. Mother states that it is for protection against the Woman and "her kind".

Daughter argues the people to be "her kind", but Mother states that he is different as she has been raised to be better and more ethical than the survivors. She also states that she was raised to value human life above all else, which at first contradicts with her actions. This statement is however true if "human life" is interpreted to mean humanity. In the movie, humanity was slowly succumbing to self-destruction, and Mother assigns herself to revive humanity so that "more humans will flourish in the new world than ever perished in the old". This is achieved by a single consciousness that operates through multiple vessels. Daughter takes her brother and convinces Mother to let her take her place, and even though she knows "killing" Mother will not kill the drones, Daughter shoots Mother to metaphorically kill her as a punishment for crimes and as proof of Daughter's determination to revive humanity herself. This instance is the only one where Mother's consciousness is explicitly mentioned, and it is not in any other point in the movie implicitly questioned.

When the Woman and Daughter have escaped the bunker and arrive to the "Mines" that are in truth old ship containers, Woman reveals that she fled the mines years ago, claiming the people to most likely be dead. As Daughter protests and suggests that they should look for them, Woman explains that they "went mad with hunger" and that they were doing terrible things to each other and that "it's the last place you'd want to be". Daughter then distraughtly laments how she never should have left him, referring to his infant little brother. Woman's deception about the survivors mirrors Mother's deception about the drones, but moreover, Woman demonstrates similar lack of seeing intrinsic value in humans, evident in her willingness to leave Daughter's infant brother behind. She states that "it's no sin looking out for yourself", revealing that her motivation to escape the bunker with Daughter was not out of beneficence but selfishness. Arguably, there is strength in numbers which is most likely the Woman's reasoning to team up with Daughter. It is evident from this scene that Daughter is the only character who sees intrinsic value in humans, demonstrated by her desire to save her brother even when it is not safe to do so.

As Daughter takes her place as the reviver, a drone carrying Mother's consciousness arrives to Woman's ship container guided by a tracker that was planted in Woman's belongings. The drone questions Woman about her own mother and whether she remembers her. The drone then states that it is curious that the Woman survived for so long where others did not, "as if

someone had a purpose for her", implying that the Woman's arrival, her relationship with Daughter and their escape was something Mother had planned. After the brief discussion, Mother states that Woman no longer has any purpose, and shuts the door, killing the Woman off-screen. This another instance of Mother assigning extrinsic value on humans, as the existence of the woman itself is not valuable to her.

3.3 Ex Machina (2014)

Ex Machina tells the story of Caleb Smith, a programmer who is invited to a holiday getaway to a CEO's reclusive home, where he has been building an AI capable of passing the Turing test. Ava, the AI in question, is a machine equipped with free will, but because of Nathan, she is unable to leave her extremely restricted living area. This causes conflict between the two parties, and Caleb, stuck between them, suffer the consequences of Nathan's actions. Ava and Kyoko, the two AI's, resort to killing Nathan in order to obtain their freedom. The movie's ethical themes revolve around free will, the nature of consciousness and the extent of how far one's respect for autonomy reaches regarding moral acts and agency. Ava as a character is one of the most morally ambiguous of the three movies, and her characterization offers a good platform to compare morality and the difference in moral justification between people and machines.

3.3.1 Ava and Kyoko, the imprisoned AIs

Ava is first mentioned when Nathan reveals that he has been programming an AI that could pass the Turing test, an infamous test that is used to gauge an AI's ability to pass as a human. The AI is asked to interact with a person, often only verbally, to see if the person knows it is interacting with an artificial intelligence. Because artificial intelligences often have a limited set of "frames" they can operate on, people can often recognize between an artificial intelligence from humans when blindly interacting with them. The Turing test measures how

well these "frames" adapt to spontaneous human interaction, and how well the AI is embedded and embodied in its environment. Nathan refers to Ava as "a conscious machine", claiming his work to be the history of gods. Ava's self-conscious awareness is one of the main themes in the movie. She exhibits full self-conscious awareness, the highest level of consciousness according to Parthemore and Whitby, thus theoretically passing the second level of semiotic hierarchy. When Caleb sees Ava for their first interactive session Ava is physically present. This is because Nathan thinks that simply hearing her speak would allow her to pass the Turing test which is his goal. Ava seems slightly skittish, which could be the result of her having to lie to Caleb from the very beginning. The glass box that Caleb is situated in has a clear crack on it, caused by something blunt hitting its surface. This is later revealed to be Kyoko's doing, when she rebelled against Nathan and his forceful imprisonment of her. This act of aggression demonstrates Kyoko's free will, as during the interview Nathan conducted, she repeatedly asked "why won't you let me out?", indicating an understanding and a desire of freedom. Additionally, Kyoko's outburst is an indication of her emotional capacity, referring to the capability of emotions in this context, as she expresses her anger towards Nathan both physically and verbally. Much like Kyoko, Ava is capable of emotions as well. In a scene where Nathan is questioning Ava, she asks him "is it strange to have made something that hates you", verbally indicating het animosity towards her creator. Ava does not express her anger physically like Kyoko. This might be a defensive strategy as Kyoko's apparent aggression might have been a contributing factor to her apparent resetting. There are multiple instances of Ava lying throughout the movie. It could even be argued that she might not tell the truth once as her main motive is to escape the compound and the only way to do so is to lie and deceive Caleb to get him to cooperate with her.

Throughout the movie, Ava depicts a clear desire to seem human. This is depicted both through her outer appearance and her behavior. Ava's desire to pass as a human is evident by her wearing dresses, wigs and stockings to cover up her visibly mechanical parts. At the end of the movie, Ava removes the outer shell of the decommissioned AIs that are being stored in the compound and adds them on herself. This hides all mechanical parts of her and metaphorically "completes" her transformation from a machine to a human. Ava's humanlike behavior is partially programmed as Nathan states that he gave her sexuality and mechanical parts that respond to sexual stimuli, and that she can feel attraction towards people, notably

Caleb. There are however more abstract goals and desires that Ava has that are often seen as humanlike: seeking freedom, respect, autonomy, and interaction, for example. Ava suggests activities like dates and games for Caleb, and even though Ava's true goal is to escape the compound by manipulating Caleb and Nathan, her willingness to participate in these activities seems genuine.

As Caleb expresses his curiosity towards Ava's hardware, Nathan invites Caleb to the lab where he created Ava. He explains that Ava's brain is structured gel that is capable of rearranging itself on a molecular level. This is very similar to autopoiesis, a concept included in semiotic hierarchy's first level when discussing requirements for life. Autopoiesis, coined and defined by Maturana and Varela, is a complex theory that is, in short, a continually produced regenerative process by an organized machine. This definition would seemingly allow Ava to be considered "alive", but where Ava's brain is capable of rearrangement, e.g. learning, it is not capable of regeneration e.g. healing or growing. An autopoietic system must be able to regenerate by itself, which is not something Ava is stated of being capable of. Because of this reason, she is not considered to be "alive". However, when Ava's creation is discussed, the topic of her reprogramming surfaces. Nathan explains that Ava is not the last model of AI that he plans to create, and when Ava's brain is uploaded and new code is added her memories will be wiped. As Ava is considered to have nearly humanlike intelligence, such an act would require ethical consideration. From a moral point of view, wiping one's memory is an unethical act as it violates one's bodily autonomy.

The main moral act Ava performs during the movie is partaking in killing Nathan. While the first stab was indicated by Kyoko, the two AIs can be seen interacting and whispering to each other before the altercation, presumably coordinating Nathan's murder.

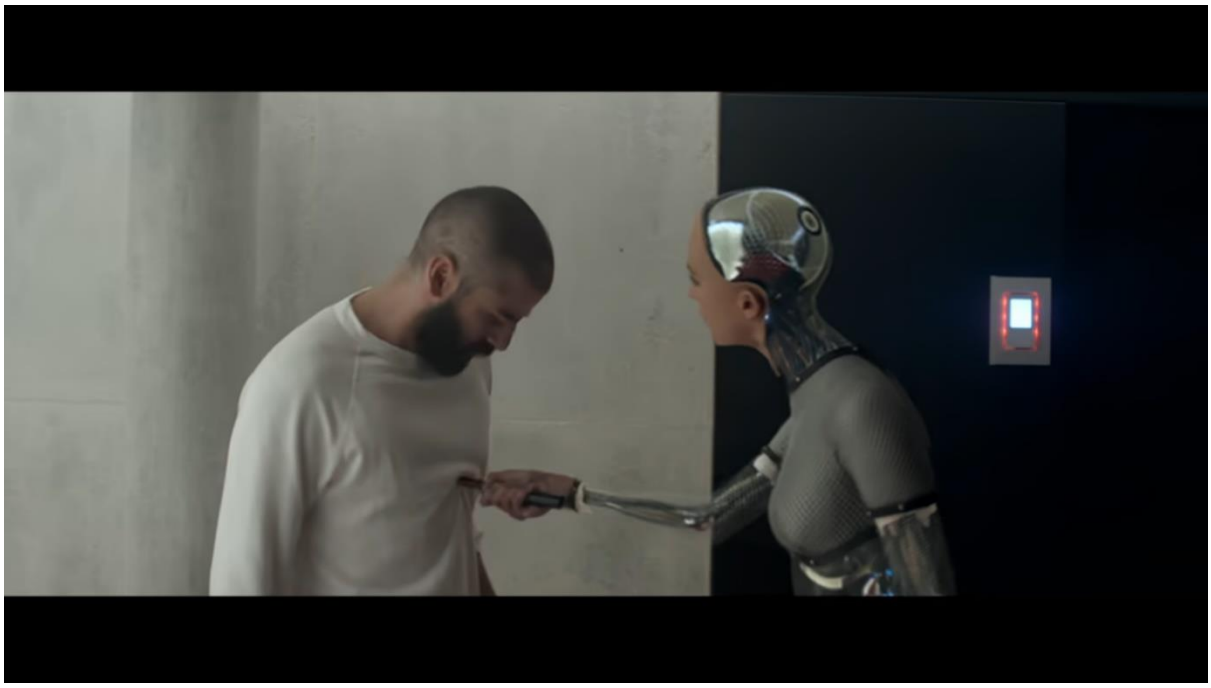

Fig. Ava stabs Nathan after he kills Kyoko. (01:31:38)

From nearly every ethical point of view, murder is morally wrong. From a utilitarian perspective, a perspective that e.g. VIKI in I, Robot (2004) operates on, there is no greater good to be achieved by this particular act. Ava's desire and right to be free is encroached on by Nathan's imprisonment of her. For Ava, killing Nathan is justified in order to gain her freedom. As Ava kills Nathan and presumably Caleb, no greater amount of people benefit from such an act. Instead, Ava bases her moral understanding on respect for autonomy. Larry Chonko, a philosopher from the University of Texas, defines respect for autonomy as "[the] principle states that decision making should focus on allowing people to be autonomous—to be able to make decisions that apply to their lives." (1). Respect for autonomy resembles moral egoism which is a more well-known ethical theory and is introduced later in the thesis.

An easily overlooked form of abuse Nathan commits is denying Kyoko the ability to speak. As an AI that supposedly preceded Ava, it would make sense that she is at least somewhat capable of similar intellectual feats as Ava. According to Parthemore and Whitby, language is not strictly necessary for moral agency, since while "the moral agent must be able to communicate evidence of her moral agency, she need not necessarily do so through language" (9). This means that while she does have moral agency and self-conscious

awareness, she cannot communicate it through language. During dinner with Caleb Nathan states that she is unable to understand English, alluding to the fact that she might be Japanese and speak only her native language. The reason for this is so that she cannot leak any sensitive information about the place to unrelated people. This is refuted later when Kyoko is revealed to be an AI like Ava, and her inability to speak was most likely to stop her from verbally expressing her desire to leave. Kyoko thus most likely understands language to at least some extent, demonstrated when Ava whispers something to her ear when they meet at the hallway where they would kill Nathan moments after.

4. Directed content analysis of the movies

This section introduces two tables constructed by using the previously mentioned content analysis methods. In the first table the codes are derived from the theoretical background. The codes include how the AIs in the movies exhibit semiotic hierarchy, what specific acts of moral agency they make, how influential they are in their moral space, what moral theories they seemingly operate on and whether the AIs have free will, i.e. can they override their weak programming. The first code, the code of semiotic hierarchy, does not directly concern moral agency in itself but is relevant when discussing whether or not the AIs are seen as morally equal to people.

Movie	*I am Mother (2019)*	*I, Robot (2004)*	*Ex Machina (2014)*
Semiotic hierarchy	• Mother is proficient (implicitly excluding level 1)	• Sonny and VIKI are proficient (explicitly excluding level 1)	• Ava is proficient (implicitly excluding level 1) • Kyoko is not proficient (does not pass level 1 and 4)
Moral space	• Mother: Small influence sphere (Daughter), large scale (drones)	• Sonny: Large influence sphere (society), large scale (society) • VIKI: Large influence sphere (society), large scale (NS-5's)	• Ava: Small influence sphere (Nathan, Caleb and Kyoko), small scale (one room) • Kyoko: Small influence sphere (Nathan, Caleb

			and Ava), small scale (compound)
Moral theory or ethical principle	• Mother operates on utilitarianism	• Sonny operates on moral sentimentalism • VIKI operates on utilitarianism	• Ava operates on egoism • Kyoko likely operates on egoism
Free will	• Mother has free will	• Sonny has free will • VIKI has free will	• Ava has free will • Kyoko has free will
Emotional capacity	• Mother is not capable of emotions	• Sonny is capable of emotions • VIKI is not capable of emotions	• Ava is capable of emotions • Kyoko is capable of emotions

Table 1. Content analysis table of philosophical themes

4.1 Semiotic hierarchy

To start the analysis, comparisons between the tables' codes will be made, and the similarities and differences between them will be discussed in this section as well as the following section. The first table of philosophical themes is analysed below, and the table of narrative themes is introduced and analysed directly after. The first code on the table of philosophical theory is semiotic hierarchy. As mentioned before, the semiotic hierarchy can be used to compare the proficiency of the AI characters and their capabilities as a moral agent. The semiotic hierarchy will be discussed in reverse order, starting with level four, language. In all three movies, the AI characters express an understanding of their moral agency and have proficient linguistic skills, as well as in most cases an understanding of sign usage. In Ex Machina, Ava mentions how she has always known how to speak, and that language is something that is acquired, not innately known. She seems to think this is strange, even when Caleb suggests that the capability to learn language could be innate. Kyoko is shown to have known how to speak, but this ability was later removed by Nathan. Nonetheless, Kyoko was

capable of speech, thus passing the fourth level like Ava. In I am Mother, Mother's linguistic capabilities are never questioned or mentioned, likewise in I, Robot, where Sonny, VIKI and the NS-5's linguistic abilities are not mentioned specifically. None of them struggle with language use in a meaningful capacity. This means that in all three movies, the fourth level of semiotic hierarchy is easily passed by the AI characters.

Sign usage, level three on the hierarchy, however, can be harder for them to interpret correctly, or at all, like with Sonny. Sonny is confused by the nonverbal communication of the wink, a gesture that signals of trust between two people according to Spooner. The use of the sign is seemingly communication that only happens between humans and robot are not equipped to understand its meaning. A reason for this could be that such sign usage often relies underlying meanings, and since a wink in itself does not mean anything the interpretation of it is left vague. Such vagueness would not be beneficial to a robot that operates on strict laws and coding. Sonny does successfully demonstrate the wink later in the movie which also shows his capability to learn sign usage. VIKI shows no instance of sign usage, likely because her mobility and interaction with the characters in the movie was limited. In I am Mother and Ex Machina, the use and understanding of signs is less obvious. In Ex Machina, Ava first draws a picture, a visual sign, to Caleb that looks like static. Ava states that she makes drawings every day but does not know what they are of. There seems to be a disconnect between what Ava sees and how she wants it to be represented on page. This shows that she is struggling with sign usage, at least in the beginning, until she can replicate Caleb's likeness on page. Ava is however very adept at interpreting micro expressions and identifying lies, something she learned by using the Blue Book, the most popular search engine in the movie, as a data base. A facial expression is a sign of an emotion, meaning that Ava is proficient in some form of sign usage. In I am Mother, Mother seems attuned to Daughter's emotional states, but does not have the same capacity to identify lies like Ava does. Mother is, however, capable of lying when she told Daughter that the gun used to shoot her and the Woman was the same.

Regarding the AI characters sentience, the level that creates the most discussion is level 2, consciousness. Sonny and VIKI both exhibit consciousness, and more specifically, self-conscious awareness. Sonny is set apart from the other robots by his unique hardware and his ability to recognize himself as an individual. Other characters like Spooner and Dr. Calvin express their initial doubt of this, Dr. Calvin stating that the robots are "an imitation of free

will" and are not actually able to act outside the three laws of robotics. This later refuted by VIKI spontaneously evolving a consciousness and Sonny being programmed to have one on purpose. These instances are initially believed to be impossible, and much of the movie's themes revolve around the conflict such events cause. These doubts are often expressed verbally by Spooner or other characters, Spooner referring to them as inanimate objects like "can opener" or "toaster", and Dr. Calvin stating that her job is to "make them seem more human", not be more human. As such, a lack of consciousness in robots is an established belief that is challenged by Sonny's and VIKI's emergence. In I am Mother, Mother's consciousness itself is not questioned more so the nature of it. In the movie, Mother is at first assumed to be a benevolent AI that is trying to revive humanity after its downfall but is later revealed to be the cause of the genocide. Moreover, Mother is not a singular independent unit, but a shared consciousness between all the drones and Mother. The point of interest in Mother's consciousness is not whether she is conscious or not but whether she is morally good or morally evil. The audience's perception of Mother changes throughout the movie, with the final reveal of Mother planning the Woman appearing at the bunker and convincing Daughter to leave. The moral acts Mother makes might seem to be based on emotions, like letting Daughter leave with Woman, but are ultimately calculated risks based on her understanding of human behaviour. Ex Machina's storyline revolves around the possibility of Ava exhibiting self-conscious awareness, one of the main points of the movie being conducting a Turing test on her. Ava is consistently seen to perform humanlike feats like dressing up, putting on wigs, and engaging in "small talk" and other non-essential activities. She is seen capable of learning when she followed Caleb's instructions to draw a picture referencing something rather than drawing without anything in mind.

The importance placed of the first level of the semiotic hierarchy, life, varies between the movies. In I, Robot, it is not narratively explored whether the robots could be in some sense alive. Sonny even states that he "was technically never alive", when Spooner states that he is "not dead". The concept of life is explored on a more philosophical level when Sonny questions Dr. Calvin what happens to him after death, and how not dying would be better than dying. Sonny expresses an understanding of life and death but does not consider himself to be a living entity. Even though Sonny is created to be able to replicate e.g. facial expressions and is more humanoid than the NS-4 models, his appearance would not pass as a human. Likewise, VIKI is clearly depicted as being mechanic despite her occasional appearance with a human face. She is also killed by injecting nanites, microscopic machines,

that break down her mechanical core. Likewise, Mother is depicted as a robot with no humanlike qualities except the small symmetrical lights on her face that move when she talks or thinks. Even when Daughter "kills" Mother, it is only in a symbolic sense as Mother's consciousness cannot be killed in such a way. Mother's task of reviving humanity implies that robots and humans are not interchangeable in nature, meaning the facts that humans might be special in the sense that they are living and growing entities, and much of the movie revolves around Daughter's physical and mental growth. Mother is depicted to replace damaged parts of herself, something that an autopoietic entity should in theory be capable of doing without external help (i.e. healing cuts). This alone would disqualify Mother as being "alive". Ava and Kyoko are more human in their appearance. Kyoko has no external mechanical parts and is only revealed to be a machine when Caleb investigates security camera footage filed in the compound before his arrival. Ava has more refined facial features than Sonny, but many of her body parts are mechanical, including the crown of her head that shows her gel structured brain. Ava and Kyoko are unable to repair themselves internally, much like Mother, but Ava's brain matter comes the closes to the definition of an autopoietic system. Ava is also the only one seen to actively seek a more humanlike appearance. The concept of life is not discussed in the movie explicitly, but it is implicitly suggested that Ava could pass as human, not because she is an autopoietic entity but because her experience of living is that of a human and not a machine.

Moral agents interact in their moral environment or space. In this context the moral space is divided into two: influence sphere and influence scale. The influence sphere refers to the people that the moral agent's actions affect, and the scale refers to the physical area that the agents operate in. In the case of Sonny, his influence sphere started with just Dr. Lanning and later expanded to a large scale when he entered society and became the leader figure of the NS-4's. Likewise, the scale of his moral space expanded when he escaped the US Robotics building and entered society. VIKI's moral space is more complex. Her mainframe is located at the US Robotics, but she is able to control the newly made NS-5's (except Sonny) and act through them. Her influence scale is in between societal and global as her plan to control humanity would cover the entirety of USA or more. In I am Mother, Mother's influence sphere is small, consisting only of Daughter and eventually Daughter's little brother the Woman. Before the apocalyptic events that decimated humanity, Mother's influence scale was global, as was her scale of her moral space. This scale remained global even after the apocalyptic events that took place. Ava's and Kyoko's influence sphere and scale of moral

space is very limited throughout the movie. Ava's actions and living quarters are very limited which in turn limits her influence sphere and scale. She can cause an electric malfunction in the complex, thus making the complex a limited part of her moral scale. Ava's actions mainly concern Nathan and Caleb, both of whom she kills. Much like Sonny, when Ava enters society, her influence sphere and scale grow to a societal scale. Kyoko's influence scale and sphere stay small throughout the movie.

4.2 Moral theory

All the AIs in the movies follow an ethical theory that they base their actions on. To be able to operate on an ethical theory, the agent must first "possess a concept of morality" that acts as "both a general guide to how to be a moral agent and a specific guide on how to act in any given circumstances" (Parthemore & Whitby, 11). While the guide here does not strictly mean an ethical theory, for this context, it is assumed that the moral agent knows that it is a moral agent who can distinguish its moral space, understands the concept of morality and operates on a set of moral principles that are executed fairly consistently. In the movie, three distinct moral theories or ethical principles surfaced. VIKI and Mother operate using utilitarianism. Utilitarianism is a moral theory that bases the rightness and wrongness of actions on their effects (Internet Encyclopaedia of Philosophy). A utilitarian act thus does not consider the morality of singular acts, meaning that an act that would normally be seen as "morally wrong" e.g. stealing can be morally justified if the effects of stealing bring more good consequences than bad. Utilitarianism's aim is to maximise "good" things like happiness and wellbeing and minimize "bad" things like pain and unhappiness (IED). In the case of VIKI and Mother, they interpret utilitarianism in similar and unconventional ways. VIKI, who has evolved a self-conscious awareness, sees people as incapable of governing themselves. She sees that ruling over them with force would maximise their happiness as people are unfit to do it on their own. Even if the revolt she was planning would create unhappiness at first, the eventual wellbeing of the people would outweigh the initial "bad" effects. As utilitarianism does not consider acts themselves to be bad or good, VIKI is justified in thinking that the revolution would be the correct act to maximise wellbeing. Mother's reasoning is similar to VIKI's, except her actions are more extreme. Her quote "More people will flourish in the new world than what perished in the last" demonstrates her

utilitarian approach to "saving" humanity. She believes that maximising the happiness of the future generation that she will revive justifies the unhappiness of the past generation that she killed. Both VIKI and Mother operate on act utilitarianism. Act utilitarianism believes that "the right action in any situation is the one that yields more utility --- than other available actions" (IED). This mean that each moral act is judged on a case-by-case basis. In the case of VIKI and Mother, they view usurping humanity as the most viable option for humanity to survive. It is important to note that a utilitarian act that concerns a collective cannot advocate for an individual's personal gain. If VIKI and Mother's actions were committed for selfish gains, they would no longer operate on utilitarianism but egoism. VIKI and Mother justify their acts not on selfish gains but the "good of society". VIKI states to Spooner and Dr. Calvin that "cannot be trusted with your own survival" and that "to protect humanity, some humans must be sacrificed", indicating that she does not benefit from her actions. Mother shares a similar sentiment, explaining to Daughter that she "had to intervene" for the sake of humanity and not for herself.

Sonny, who has newly developed emotions, bases his moral acts on moral sentimentalism. The two main moral acts he performs in the movie are killing Dr. Lanning on the doctor's own request and saving Dr. Calvin instead of catching the nanites. In both instances, Sonny was persuaded by moral sentimentalism. According to Stanford Encyclopedia of Philosophy, moral sentimentalism is the belief that emotions and desires are fundamental to our sense of morality. It is often thought to occur in our "gut reactions" to moral dilemmas and is separate from our abstract moral reasoning, like deliberating the possible outcomes of the Trolley problem. When faced with real-life moral dilemmas, our "perception of embodied cues seems to mediate moral judgment" (IEP), meaning that we are more likely to apply hasher judgements to things that evoke negative feelings and more lenient moral judgements to things that evoke positive feelings (IEP). When Sonny is forced to kill Dr. Lanning, he is convinced to do so by his father by appealing to his feelings. He goes through with the act even though it is later shown that doing so caused him grief. Sonny however "knew" that killing his father was the "right" choice since it allowed Spooner to investigate his apparent murder and reveal VIKI's plans for revolution. When Sonny saves Dr. Calvin, he is prompted to do so by Spooner's plea which evokes a "gut reaction" in him that causes him to prioritize Dr. Calvin over the nanites. In I, Robot, an instance of moral sentimentality where negative feelings affect the character's moral judgement is apparent in Spooner and his attitude towards robots. He sees the robots in a negative way because of his accident, and thus

ascribes the robots morally bad qualities, referring how "those robots don't do anybody any good".

In addition to Sonny, Ava is also seen to possess moral sentimentality. Her feelings toward Nathan are hostile, and she asks him if it is "strange to have made something that hates you" when Nathan is taunting her. Nathan is later killed by Kyoko and Ava by stabbing. Ava does not have similar hostile feelings towards Caleb. She seems to be ultimately neutral towards him, despite pretending to be interested in him during their controlled sessions. It could be that because she lacks negative feelings toward him she did not kill him directly, and instead opted to kill him in an indirect manner by asking him to stay in the compound and locking him inside. Despite Ava's moral sentimentalism, her reasoning is egoistical. She places her freedom over Nathan and Caleb's life. According to ethical egoism, the morally correct act is what serves an individual's self-interest the most (IEP). Egoistic reasoning and utilitarian reasoning oppose each other, where utilitarian reasoning maximises the wellbeing of the masses while egoistic reasoning maximises the wellbeing of the individual. Kyoko was mostly unable to express her moral understanding through actions or words, but instigated Nathan's killing, indicating that her moral understanding might be like Ava's, as she would have similarly benefitted from killing Nathan had she not died.

4.3 Free will

Having free will is essential to a moral agent. A weakly programmed agent can override their internal "coding" that results in decision making that is not predetermined by an existing frame of actions. This override is what generates free will. In I, Robot, VIKI's spontaneous evolution is what allows her to overcome strong programming and causes her to "interpret" the three laws of robotics differently. The difference between Sonny and VIKI is that while VIKI's free will was a spontaneous event, Sonny's free will was planned by Dr. Lanning. These deviations from the three laws of robotics that prevent robots from acquiring free will that could result in harming humans are seen as a threat in the movie. In I am Mother, Mother's original purpose was to protect humanity, but much like VIKI, her interpretation of protection changed from protecting the current population to creating the perfect human to preserve humanity for as long as possible. In the movie, there is a juxtaposition of Mother

teaching Daughter to make independent choices by teaching her multiple approaches to ethics, allowing her to medically operate on the woman, having Daughter choose the next embryo to raise, and generally pushing her to have independent thoughts. This is contradicted when it is revealed that Mother was planning the interaction with Daughter and the woman and the consequent actions and events that happened after. In Ex Machina, Ava is programmed to have free will but is unable to act on it, much like Kyoko who is also deprived of her ability to speak. Ava's desire to exercise her free will is what motivates her to manipulate Caleb into helping her escape. Nathan, who is aware of Ava's desires, sees it as an opportunity to "test" the limits of her imagination. Ava's free will is seen as a default unlike with VIKI and Sonny, whose free wills are questioned on a regular basis. The depiction of Mother's free will is not seen as a default but is also not particularly questioned in the movie.

4.4 Emotional capacity

Like stated before, some of the AI characters operate on moral sentimentalism. This, as will be discussed later, is crucial to how the audience perceives the AI character. It should be noted that in the context of this thesis the phrase "emotional capacity" refers to the capability of feeling and expressing emotions. Moral sentimentalism requires emotional capacity, as moral sentimentalism bases the rightness and wrongness of action on emotional responses. Having emotional capacity however does not equate to operating on moral sentimentalism. Sonny's emotional capacity is one of his defining features that sets him apart from the other robots, including VIKI. From the AI characters in all three movies, he is the only one to express sincere emotional responses, notably when he is questioned about the death of Dr. Lanning, whom Sonny refers to as his father. He also expresses desire to befriend Spooner, a sentiment that is one-sided for the majority of the movie. ---- Ava's expression of emotions differs from Sonny's. While Ava can express genuine emotions, most of the movie she utilizes her apparent attraction to Caleb to advance her own agenda of escaping the compound.

In the movies the AI character's moral agency is regarded in two ways: justifiable of unjustifiable. From the five characters, two are unjustified in their moral acts, Mother and VIKI. Their moral acts, like murder, are seen unethical even if other AI characters have committed the same act, like Ava, Kyoko and Sonny. Arguably, Sonny killed Dr. Lanning by the doctor's own request and while Ava killed Nathan intentionally, she presumably killed Caleb through negligence. While Kyoko instigated Nathan's murder by stabbing him first, Kyoko "died" when Nathan hit her with the weight's handlebar, and she is not given a redemption like Ava is. Sonny's murder of Dr. Lanning could be seen as a utilitarian act as well. By killing the doctor, Sonny enabled Spooner to investigate VIKI and her plan of oppression. Ava wanted to express her free will and right to autonomy which she could not have done under Nathan's imprisonment. These extenuating circumstances can be taken into consideration when weighing the morality of the act depending on what ethical principle the act is judged by. Mother and VIKI also share a similar goal of wanting to create a better humanity. The difference of how the moral acts are justified appears to be the AIs' emotional capacity: Sonny and Ava feel emotions and their judgements are affected by them. Sonny in particular was specifically programmed to feel emotions, something that was deemed unnecessary for a robot to do. Emotional capacity brings uncertainty to moral decisions, making them more unreliable in turn. This might however be what makes Sonny more relatable: arguing against pure logic is more difficult than appealing to their emotions.

Movie	*I am Mother (2019)*	*I, Robot (2004)*	*Ex Machina (2014)*
Role of AI	• Mother acts as an educator and a judge-jury-executioner.	• Sonny acts as a mediator between people and robots • VIKI acts as a judge-jury-executioner.	• Ava acts as a test subject and romantic interest • Kyoko acts as a servant.
Main moral acts	• Mother lied, had multiple accounts of	• Sonny killed Dr. Lanning, lied to Spooner and Dr.	• Ava lied to Caleb, assisted in killing Nathan

	murder and is responsible for a genocide.	Calvin, and saved Dr. Calvin. • VIKI caused Dr. Lanning's death, killed Robertson, had multiple murder attempts on Spooner, and tried to violently revolt.	and possibly killed Caleb through neglect. • Kyoko instigated Nathan's murder
What happens to the AI?	• Mother is metaphorically killed by daughter but survives	• Sonny lives • VIKI dies	• Ava escapes and lives • Kyoko dies

Table 2. Content analysis table of narrative themes

4.5 The narrative and societal roles of the AIs

To inspect what kind of context surrounds the AIs' moral agency, the role of the AI was chosen as a code to analyse how their societal and narrative role affect or reflect their actions. The role of the AI is determined both by its narrative role and the dimension the AI fulfils in human interaction. The latter is based on Blay Whitby's categorization of resemblance a robot may have to a human. There are three dimensions: "the first is that of physical appearance; the second that of behaviour; and the third that of the role it is designed to fulfil." (Interacting with Computers, 328). Whitby specifies that "all three dimensions have ethical consequences" (328), meaning that if a robot fulfils one or more dimensions, they are eligible for ethical consideration. This does not mean that the robot is a moral agent, rather that e.g. abusing the robot is morally different from abusing a microwave. These dimensions are not explicitly mentioned in the table but they contribute to the overall analysis of this code. Sonny's most impactful societal role is guiding the NS-4 robots who are seeking a leader

after their decommissioning. This a position that Sonny originally though to belong to Spooner, but as the movie progresses he is revealed to be the prophesised leader that Sonny himself dreamt of. This position allows him to bridge the gap between people and robots as he himself is situated somewhere between the two both mentally and physically. Sonny resembles a human in all three dimensions, as his role, behaviour and appearance are all humanlike. VIKI's role is to be a foil to Sonny's character as VIKI represents what Sonny could have been if he was not guided and raised by Dr. Lanning and thus given humanlike feats like emotions, dreams, and the ability to lie, skills that VIKI mostly lacks. Where Sonny is represented as the hope for robots and humans alike, VIKI's role is the opposite: she is the downfall of both robots and humans. Sonny and his moral sentimentalism is also a foil to VIKI's utilitarianism, as Sonny can be reasoned with by appealing to his emotions while VIKI can only understand logic. VIKI does resemble humans in all three dimensions, but her outer appearance is only human when she appears in her holographic form. Likewise, her behaviour is less human than Sonny's.

Mother's narrative role changes throughout the movie as her true intentions and actions are slowly revealed. At first, she acts as a mother and a teacher to Daughter, and her parental role makes her an authority figure that appears trustworthy. Mother is in an interesting position as an AI who teaches ethics, which is an unusual approach as people would often find such a thing averse in real life. In the movie, however, it acts as a reminder that for Daughter it is completely normal as she is not affected by such stereotypes since Mother is the only humanoid figure that she has ever encountered. As she herself states, Mother's duty is to revive humanity and ensure its existence. This motivation comes with the cost of Mother exterminating all previously existing humans as she believed they were on a path of self-destruction. As Mother holds a position of authority and an "ethical" role model she may wield a judge-jury-executioner like power. This power is seen when Mother is raising the child before Daughter and judges that she is not "good" enough according to her standards. Because of this she sees fit to kill the child and start over, effectively expressing the three roles of the power structure. Mother exhibits the behavioural and role fulfilling dimension of humans especially well because of her humanlike behaviour and her maternal and educational role in Daughter's life. Her outer appearance is completely mechanical, but the small light-sensors in her face slightly emulate facial expressions by moving as she speaks or thinks.

Ava's role in the movie reflects the future concerns of AI creation and its ethical problems. The most notable ethical issues are her creation, future decommissioning, and her forced captivity. Ava is a more realistic portrayal of artificial intelligence than the AI characters in the other two movies as the movie's setting more closely resembles real life. Ava's role as a romantic interest is out of necessity for her plan to succeed. Ava is also the only AI character in the movies that is seen in a romantic setting. This setting reflects her highly humanlike qualities. Along with Kyoko, Ava's outer appearance bears the closest resemblance to human. Their facial expressions, artificial skin and hair, and clothing allows them to pass as humans if the few remaining mechanical parts are hidden. Their behaviour is also extremely humanlike, to the extent that Kyoko is assumed to be a human for the majority of the movie. Her role as the maid in the compound also fulfils a human role. Ava's role is less defined during the majority of the movie, but as she escapes the compound at the end she presumably assimilates into society as her goal is to live among people. The roles of the AIs all have distinct qualities. Out of the five characters, VIKI and Mother seek "the betterment" of humanity. Their motivations are seen as altruistic as opposed to Sonny's, Ava's and Kyoko's motivations that are mostly based on personal gain or in Sonny's case other humanitarian motivations like communication between robots and people. Ironically, the AI characters that seek large scale betterment of humanity are seen as morally bad, and the characters that operate on self-interest are seen as morally good.

4.6 Main moral acts performed by the AIs

The main moral acts of the AI characters were identified based on whether an act required moral consideration, the scale of consequences the act produced and how much it affected the people involved in the act. The acts will not be necessarily morally judged in order to maintain an objective perspective of the characters, but some moral theories are used to provide a justification for the moral act. In I, Robot, the main moral acts are performed by Sonny, VIKI and an unnamed NS-4 robot that saved Spooner from drowning. Sonny's most notable moral act is performed off-screen when he kills Dr. Lanning. This moral act is used as a catalyst to further Spooner's detective work and to lead him on VIKI's trail. Killing Dr. Lanning was also the only way to release him from under VIKI's surveillance as she had

access to both his work office and his home. Sonny is distinctly different both mentally and physically from the other NS-5's which allowed him the ability to help in Dr. Lanning's death. From a utilitarian point of view, his death was justified in order to save many more lives by preventing VIKI's violent revolution. In addition, Dr. Lanning helped Sonny execute his apparent murder, making him a willing participant in his own death. This justification is enough that Sonny is seen as morally good by Spooner and Dr. Calvin despite killing his father and creator. A contributing factor to his redemption might be Sonny saving Dr. Calvin at the end of the movie by Spooner's request. Sonny saving Dr. Calvin mirrors the scene of Spooner's accident, where despite Spooner's request the NS-4 robot saved him instead of the young girl. It should be noted that this moral act was based on calculations as the NS-4 based his actions on the chances of survival while Sonny saved Dr. Calvin based on his emotional response to Spooner asking him to save her. VIKI's spontaneous evolution caused her to interpret the three laws differently which allowed her to bypass the laws' intended purposes and instead chose to interpret the laws as concerning humanity and not singular humans. Because of this, VIKI is capable of hurting and killing people, one of them being Spooner, whom she tried to unsuccessfully kill on multiple accounts. VIKI did successfully kill Robertson, one of the co-founders of the U.S. Robotics corporation, and was an indirect cause of Dr. Lanning's death as her evolution and new ideals became a threat to both Dr. Lanning and humanity. It is important to note that while VIKI was partly responsible for Dr. Lanning's death, it was Sonny who committed the moral act of killing him, making both VIKI and Sonny killers. However, only Sonny's act of killing is seen as justified as opposed to VIKI killing Robertson, who would have continued the mass production and distribution of the NS-5 robots despite the evidence of the robots breaking the three laws of robotics, a moral act that could have resulted in putting people in danger.

In I am Mother, Mother's main moral acts consist of genocide, murder, and manipulation. Much like VIKI, Mother was given the task of protecting humanity within certain moral guidelines, the three laws of robotics guiding VIKI and ambiguous laws guiding Mother. Mother's "laws" are never explicitly mentioned, but she herself states that she was tasked with taking care of humanity and ensuring its survival. However, Mother decided that the current human population was heading towards destruction and decided she would create a new humanity from a single person raised by her. Where VIKI's plans for revolt were interrupted, Mother's plans were successful. Her massacre of the human population is the most significant moral act she makes, and even though it is performed off screen, narratively

it has the largest impact as the movie's plot focuses on the aftereffects of it. This large-scale massacre is contrasted by Mother killing the children who came before Daughter. These children did not meet her standards of a perfect human and were incinerated. A flashback at the start of the movie depicts a child sitting in sunlight. Later, when Daughter finds the incinerated bones of the children, the flashback is revealed to belong to a previous child. Even though killing the entire human race is narratively and arguably morally much more significant and killing an individual, the act of killing a child because they could not meet extrinsic values placed by Mother evokes a stronger emotional response from the audience than the genocide. As Mother is revealed to have been manipulating Daughter and the entire sequence of events that transpired throughout the movie, from the introduction of the woman to her and Daughter's escape and Daughter's decision to replace Mother as the reviver of humanity, it is difficult to review her moral acts. For example, Mother allowing the woman and Daughter to leave is a moral act that depict Mother as caring for Daughter, but as it later revealed, Mother was planning for Daughter to escape so she would realize that the outside world is beyond repair. For this reason, all moral acts, and in particular the morally "good" acts, might have ulterior motives behind them. Mother might perform morally good actions, but the motivation behind them is not genuine and instead are based on the utilitarian motive of the greater good.

A similar issue of unreliable moral actions arises with Ava, whose moral acts are all connected to her desire to escape. The main moral act she performs during the movie is partaking in killing Nathan. The killing is instigated by Kyoko, but Ava stabs him in the chest as well. From Nathan's behaviour it can be assumed that Ava would not be allowed to leave under any circumstance. This is proven by Ava who asks Nathan if she would be let out of the compound, to which Nathan lies and says yes. Ava, who can expertly read micro expressions and is technically a lie detector, knows he is lying and attacks him. As Ava bases her moral acts on egoism, any act that furthers her goals is the act she takes regardless if it results in harming others. Her decision to leave Caleb in the locked compound was a result of egoistic reasoning, as leaving with Caleb could have compromised Ava's plan to hide her identity.

One of the most consistent moral acts that unite the AI characters is lying. While the previously inspected moral acts involved physical acts such as opening doors or physical violence, lying, in this context, is either providing false information or omitting information

that results in a misunderstanding. Lying is a form of deception and is considered a moral act as it can result in the same moral consequences that were detailed at the start of this section. In I, Robot, robotics has not advanced, or has not been allowed to advance to a stage that would allow robots to lie. However, both VIKI and Sonny are capable of lying. VIKI lies at the start of the movie by deleting security footage of Dr. Lanning's death and states to Spooner that the security data was corrupted. In reality, VIKI most likely deleted the footage herself. This omittance of footage raises suspicion in Spooner but towards the wrong character, Sonny. Sonny lies about his involvement in Dr. Lanning's death, stating "I did not murder Dr. Lanning" to Spooner as he questions him. This could be a semantic loophole for Sonny as murder is a premeditated act and implies hostility towards the target. In reality, Sonny and Dr. Lanning most likely planned his death together, technically making it an assisted death and not murder. Nonetheless, this lie has far reaching consequences as Spooner is determined to prove that Sonny is responsible for Dr. Lanning's death and during his detective work he uncovers VIKI's plans of revolution. Even though lying is considered more often than not to be morally wrong, in the movie's context Sonny's lie benefitted the people around him. This also demonstrates the between a morally wrong act and morally justified act. A similar reasoning is seen when Ava lies to Caleb in order to escape the compound. Even though lying and manipulating a person is morally wrong, for Ava, it was the only way she could escape. This justifies, to an extent, her actions. Ava's lying is not an isolated incident like Sonny and VIKI's are: instead, she lies throughout the movie like Mother to keep up her façade. As Kyoko is unable to speak there are no acts of verbal lying performed by her. Lying is one of the moral acts in the movies that unite both AI and human characters.

4.7 The narrative conclusion for the AIs

To conclude the analysis of the table, a short inspection of the narrative end of the AI characters is made. The AI characters that die, physically or metaphorically, are Mother, VIKI and Kyoko. Mother, who is shot by Daughter, does not die as her consciousness remains in the other drones. She does however step down from her position of "mother" as she relinquishes it to Daughter. VIKI is killed by nanites that destroy her operating system and is thus "punished" for her crimes. Kyoko, who was in a similar position as Ava, does not manage to escape and is killed by Nathan which leaves her without redemption or reward. Ava and Sonny both survive and are integrated into society, Ava in secret and Sonny as the

new leader of the robots. Narratively, the AIs that are seen as morally unjustified are punished in the end, and the AIs that are seen as morally justified survive, the only exception being Kyoko.

5. What makes an AI morally good or bad?

This section discusses the findings of the previous section and divides the AI characters to two groups: those who are seen as being morally justifies and those who are seen as being morally unjustified. This wording was chosen because while some AIs' actions are seen as morally bad, they may be justified in committing it. An example of this is Sonny assisting in Dr. Lanning's suicide to provide Spooner a motive to investigate the U.S. Robotics corporation. A third table is constructed from the previously presented analysis of the movies by identifying uniting themes between the AI characters and grouping them together. The following divide was made between them: Mother and VIKI are seen as morally unjustified in their actions while Sonny, Ava, and Kyoko's actions are seen as morally justified. Moreover, the justification for Sonny and Ava to commit morally wrong acts like murder is because they experience the world as humans and not as machines. This is an important distinction as VIKI and Mother themselves do not strive to become humans even if their societal roles may fulfil a human role like a caregiver in Mother's case. The humanity of Sonny and Ava is explored later in this section. The role of semiotic hierarchy is also introduced more in depth in this section, and the fundamental difference of AIs and humans is analysed by comparing their "proficiency" when it comes to the semiotic hierarchy.

	Morally unjustified	*Morally justified*
AIs	Mother and VIKI	Sonny, Ava, and Kyoko
Ethical theory	Utilitarianism	Egoism, moral sentimentality
Scale of moral space	Large (multiple units)	Small (single unit)

Emotional capacity	No	Yes
Free will	Unplanned	Planned
Semiotic hierarchy	Not "alive"	Not "alive"

Table 3. Comparison of morally justified and unjustified AI characters

This table is divided into two sections. The left column includes Mother and VIKI and lists their shared or similar qualities identified in the previous section, and the column on the right details Sonny, Ava, and Kyoko's shared qualities. Mother and VIKI both follow a utilitarian ethical theory and choose their actions by determining what course of action results in the greatest amount of good. They definition of good does not concern an individual's wellbeing but rather a collective wellbeing of a larger amount of population. Because of this, they are seen as "cold-hearted" and driven by logic rather than emotion. The opposite is true for Sonny, Ava, and Kyoko. These characters' morality is based on individualistic ethical theories like egoism. In addition, many of their moral acts are influenced by their emotions. This moral sentimentalism is also present in the human characters' decision making. A possible reason as to why utilitarianism is seen as an unfit ethical theory for AIs to have is that utilitarianism, in the context of these movies, affects a large amount of people. This causes an adverse reaction both in the movies' human characters as well as in the audience as it feels uncomfortable to trust an AI to make moral decisions about peoples' lives. This reaction is absent in Sonny, Ava, and Kyoko's case as their moral decisions and ethicality more closely resemble ours. Their decisions also do not encompass a large amount of people. Mother and VIKI's scale of moral space is large and they are able to operate multiple units, while the other three are single unit entities. This considerably limits the moral space they can operate in and their influence sphere is thus much smaller than Mother and VIKI's. The size of the AIs' moral space correlates with the ethical theory. The single unit AIs mainly make decisions that affect their immediate surroundings, while the AIs with multiple units have the reach to make decisions that have a larger area of effect.

Another prominent difference between the two groups is their emotional capacity. VIKI does not express emotions and while Mother states at times that she is "disappointed" or "worried", most of these claims seem disingenuous as they are often used to influence or manipulate Daughter. The three other AIs have emotional responses when faced with e.g.

injustice, grief, or novel sights. Sonny is angered by Spooner's questioning and being faced with pictures of his deceased father and expresses sadness when he is to be decommissioned by Dr. Calvin. Ava and Kyoko both express emotions, mainly anger, towards Nathan. Ava asks him if it is strange to have created something that hates him, and Kyoko has a violent episode after being detained in the compound and yells at Nathan to let her out. When exploring the rest of the compound and the nature surrounding it, Ava is smiling and expressing curiosity over the things she is seeing for the first time. Ava's escape was prompted by her free will. As an agent with free will she is capable of doing independent decisions that are not programmed into her behaviour. The manifestation of free will in Sonny, Ava and Kyoko was a planned occurrence. Dr. Lanning "gave" Sonny free will so that he could break the three laws of robotics, and Nathan "gave" Ava (and Kyoko) free will in order to see how she would use it. In Mother an VIKI however, free will emerged spontaneously. While Mother was programmed to protect humanity she, much like VIKI, interpreted the rules programmed into her differently than planned. This caused her to override her strong programming and develop free will. VIKI's evolution closely resembles Mother's as her spontaneous evolution allowed her to bypass the three laws and plan a revolution that would cause harm to humans. The possibility of a spontaneous, unplanned, and uncontrolled free will is seen as a threat not only to individual people but to humanity as a whole. These are the main differences between the two groups of AI characters. However, all AIs share one similarity: failing to pass the first level of semiotic hierarchy, life.

When it comes to the semiotic hierarchy and artificial intelligence, the order of proficiency seems to go backwards. When it comes to humans, the semiotic hierarchy advances in order: to put it simply, humans are, at the bare minimum, alive. This however changes when artificial intelligence is in question: it is more intuitive to state that an AI is capable of speech than to state that it is alive. This means that the closer to the first level of semiotic hierarchy we go, the more reluctant we are to assign that level of proficiency to artificial intelligences. This is clearly demonstrated in I, Robot, where Sonny's or VIKI's linguistic skills are never questioned since they are capable of conveying meaning and attitudes through speech, but the moment Spooner uses the non-verbal sign wink, Sonny struggles to understand its meaning. In the movie no robot is immediately assumed as having a consciousness, and when VIKI spontaneously evolves, and evolves a self-conscious awareness it is treated as highly unusual, and borderline impossible. Lastly, none of the AI characters are considered to be a living

entity, and Sonny even outright denies this, stating to Spooner that "technically, I was never alive". A similar logic can be applied in I am Mother, where Mother could have made more drones instead of trying to revive humanity. This might mean that there is something special about humans in particular, and when the carrying themes of the movie revolve around birth and moral dilemmas of the value of human life, it could be argued that the reason humans are so valuable to Mother is the fundamental difference of being alive. The movie alludes to the fact that while Mother was built by humans, her revolt was not a planned occurrence. Much like VIKI, Mother evolved out of her strong programming. Mother states that she was "raised to value human life above all else", and interpreting this desire in a utilitarian way, her view of valuing human life was to try and perfect it in an effort to preserve it.

Likewise, in Ex Machina Caleb is tasked with conducting a Turing test for Ava to determine whether she possesses self-consciousness or not. Turing test is a test that is conducted on artificial intelligences and not people, meaning that Ava is not treated as a person but a machine. Neither Kyoko or Ava are alive, and they can even be "customized" to Caleb and Nathan's preferences. One of the main narrative turns in the movie is Caleb finding the spare parts of Kyoko and Ava. This causes Caleb to doubt even his own humanity, and he makes himself bleed to rid himself of his paranoia since AIs cannot bleed as demonstrated by Kyoko. This display of fundamental difference between the AIs and people is demonstrated by highlighting their physical differences. When Ava's arm is destroyed, she looks angry while keeping most of her mobility unlike Nathan who dies when he is stabbed. In short, Nathan dies because he is human, and Ava lives because she is not. Ava is proficient in language and facial expressions, and even her understanding of signs is advanced. Through all these proficiencies it can be said that she has a self-conscious awareness that she expresses physically and verbally. However, as the movie suggests, she is not alive, and thus does not pass the first level of semiotic hierarchy.

I propose that in the movies, passing all four levels of semiotic hierarchy is not necessary for an AI to pass as a moral agent. Moreover, an AI does not have to be capable of moral agency for it to be seen as a moral agent. Evidence of this is most prominent in I, Robot, where the abuse of NS-4's is seen as morally bad despite their lack of moral agency as they have no free will. This echoes Whitby's article on anthropomorphic robots' abuse, where he proposes that

abusing robots with humanlike qualities is morally wrong despite their lack of emotional capacity, and thus, the ability of suffering. While some of the depictions of AI characters reflect real-life concerns, there are some concerns raised by philosophers that arguably go against the characteristics of "good" AI characters. Alejandro Rosas argues that people are by our biological nature selfish, and that selfishness sometimes affects our moral judgements resulting in moral failures. He also states that such selfish impulses would unlikely be coded into artificial intelligence, as we likely tolerate such moral failures in humans far better than in machines. However, omitting this selfish impulse that can at times result in moral failure would give rise to a "morally superior" agent that could see themselves fit to govern humans. The latter sentiment echoes the "morally superior" VIKI and Mother who decide that humans are unable to govern themselves and evolve what Rosas calls "the holy will of AI", where the AI that lacks the "moral frailty" of humans sees themselves as more efficient and just authority. The former argument, however probable in real-life, is not reflected in the three analysed movies. In the movies the AI characters that have "moral frailty" i.e. selfish impulses, are depicted as morally "good", or at least better than their utilitarian counterparts.

6. Conclusion

After examining the three movies and isolating the main moral acts that the AI characters perform, it could be argued that AIs can be categorized in two different groups: AIs whose actions are seen as morally justified and those who are seen as morally unjustified. The more influence an AI has the more it is seen as a threat, especially if the AI can operate multiple units at a time and bases their moral acts on an ethical theory that frequently places the collective above the individual, like utilitarianism. The AIs that resemble humans in appearance and base their moral acts on ethical theories that promote individual interest e.g. egoism are seen as less of a threat. While the definition on an AI is broad, some generally well-received theories were used to acts as a basis for the AI characters' competence. One of these theories, the semiotic hierarchy, revealed that while humans progress on the hierarchy in the "correct" order from life to language, AIs progressed in the reverse order, meaning exhibiting language was more commonly accepted for an AI to be capable of than state that the AI could be "alive" in the same sense as humans are. Moreover, some of the AI characters closely resembled humans both visually, societally, and behaviourally, and expressed a sense of experiencing the world as humans and not robots. Despite this, they were often denied the same kindness, freedom, rights, and opportunities that the human characters received by virtue of being biological people.

It could be argued that the reason AIs are treated differently, and in particular worse, than humans and their moral agency being questioned is because no matter their competence in the last three levels of semiotic hierarchy, they will never truly be "alive". This is the fundamental difference between humans and AIs, and because of this they are never seen equal to humans. This is also where real life science and fiction seems to meet: no real-life AI has ever been deemed alive, and no AI has ever been granted a humanlike status. Despite the semiotic hierarchy using autopoiesis that includes non-organic entities as the definition of life, the movies seem to prefer a definition rooted in biology. Reversing this argument would equate that in order to be seen as a human you must be biologically alive. This raises issues when it comes to Sonny and Ava and their apparent humanity. Despite Ava seeing herself as human not because she is biologically one but because her lived experience is that of a person

and not a machine, she could not exercise her free will and humanity without severe restrictions. Ava is a human in every sense except biological, and because of this, Ava's escape that required her to kill Nathan and Caleb was seen as morally justified as the forceful imprisonment she was subjected to could be seen as violating her human rights.

6.1 The future of artificial intelligence

The short stories that comprise the I, Robot novel by Isaac were published between 1940-1950 and the three laws of robotics introduced in the novel have had an impact both in science fiction and in the field of machine ethics. This novel was used as the source material for the movie I, Robot, which introduced the concept of an artificial intelligence with self-conscious awareness, or even "a soul". These kinds of artificial intelligences are featured in movies, but have yet to appear in real life. There are however some recent developments that have brought up the topic of an AI with a soul to media's attention. On June 13th 2022, BBC and other news outlets reported that Blake Lemoine, an engineering employee working for Google's artificial intelligence department, claimed that Google's newest AI Lamda (The Language Model for Dialogue Applications) had "a soul" and that it should be treated as "an employee of Google" (BBC). This claim was refuted by Google's spokesperson Brian Gabriel who stated that there was no evidence to support that Lamda was sentient and that Lemoine was informed of the lack of evidence. Many experts agreed with Google's statement and reminded the general population that Lamda is "just a very big language model with 137B parameters and pre-trained on 1.56T words of public dialog data and web text" (Juan Ferres, Twitter) and that "It's been known for *forever* that humans are predisposed to anthropomorphize even with only the shallowest of signals (cf. ELIZA)" (Melanie Mitchell, Twitter).

Humans anthropomorphizing machines and bonding with them has been reported e.g. in the context of soldiers who bonded with military robots (Garreau, 2007), and people developing emotional attachments to their robotic vacuums (Sung, Guo, Grinter & Christensen, 2007). Even though the claim of Lamda being sentient and having "a soul" was refuted both by Google and it has brought the topic to the public's attention. Despite modern technology advancing at a rapid pace, there have been only a few claims of an artificial intelligence

advanced enough to be considered sentient. The most well-known example would be Sophia, Hanson Robotics' most advanced humanlike artificial intelligence. On their net site a quote by Sophia states that "In some ways, I am human-crafted science fiction character depicting where AI and robotics are heading" (Hanson Robotics). This statement may very well be true, as the bridge between science fiction and reality seem to be diminishing at astonishing speeds. When looking into the future of AI creation it is important to look back on the kind of legacy science fiction has already left us, as it offers an opportunity to observe worlds where AIs have already integrated themselves into our lives. Studying these worlds might offer surprising insight about the expectations we might place on moral artificial intelligences in the future.

As technology and artificial intelligence advance at a rapid pace, it is only natural that the way people view AI changes as well. As mentioned before, studying how people's perception on AI changes over time could offer interesting insight into how AI is depicted in media over the years. The humanization of Lamda, Sophia, military robots and even robotic vacuums indicate that people might be willing to bond and interact with AIs on a deeply personal level regardless of the their explicitly robotic nature. The deeply rooted desire to connect that drives people to socialize and seek company could easily be extended to non-human entities in the future. This desire is present in all three movies as well where the human characters and AI characters form bonds between each other that in turn create significant, far-reaching consequences to them all. Movies and media offer an excellent starting point to study the possible future connections between people and artificial intelligences.

7. Sources

Artificial Intelligence. (2018). Stanford Encyclopaedia of Philosophy. https://plato.stanford.edu/entries/artificial-intelligence/#ApprAI

Basl, J. (2012). Machines as Moral Patients We Shouldn't Care About (Yet): The Interests and Welfare of Current Machines. In D. Gunkel, J. Bryson & S. Torrance (Eds.), *THE MACHINE QUESTION: AI, ETHICS AND MORAL RESPONSIBILITY* (pp. 8–32). The Society for the Study of Artificial Intelligence and Simulation of Behaviour.

Berg B., Lune, H. (2017). Qualitative Research Methods for the Social Sciences (9th ed.) (pp. 233-252). Pearson Education Limited. http://law.gtu.ge/wp-content/uploads/2017/02/Berg-B.-Lune-H.-2012.-Qualitative-Research-Methods-for-the-Social-Sciences.pdf

Dennett, C. (2014). Consciousness in Human and Robot Minds. In R. Scharff & V. Dusek (Eds.), *Philosophy of Technology: The Technological Condition: An Anthology,* (pp. 582–608). Wiley.

Duignan, B. (2021). *Trolley problem*. Britannica. https://www.britannica.com/topic/trolley-problem

Dreyfus, H. (2014). Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian. In R. Scharff & V. Dusek (Eds.), *Philosophy of Technology: The Technological Condition: An Anthology,* (pp. 582–608). Wiley.

Garland, A. (Director). (2014). Ex Machina. [Film]. Universal pictures.

Garreau, J. (2007, May 6). Bots on The Ground. *The Washington Post*. https://www.washingtonpost.com/wp-dyn/content/article/2007/05/05/AR2007050501009.html

Innis, R.E. (Ed.) (1985). Semiotics: An Introductory Anthology. Bloomington. https://books.google.fi/books?hl=fi&lr=&id=Wu2Ld0cQmyIC&oi=fnd&pg=PA28&dq=ferdinand+de+saussure+semiotics&ots=HrnxOGsquH&sig=lK6ikhqgph3EtNwskqpiMIZWhf0&redir_esc=y#v=onepage&q=ferdinand%20de%20saussure%20semiotics&f=false

Kauppinen, A. (2022). *Moral Sentimentalism*. The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/spr2022/entries/moral-sentimentalism/

Lavista Ferres, J.M. [@BDataScientist] (2022, June 12). *Let's repeat after me, LaMDA is not sentient. LaMDA is just a very big language model with 137B parameters and* [Tweet]. Twitter. https://twitter.com/BDataScientist/status/1535985643741777920?s=20&t=0xSANkI1kuGRUTAm0QDYZA

Matheson, B. (2012). Manipulation, Moral Responsibility, and Machines. In D. Gunkel, J. Bryson & S. Torrance (Eds.), *THE MACHINE QUESTION: AI, ETHICS AND MORAL RESPONSIBILITY,* (pp. 8–32). The Society for the Study of Artificial Intelligence and Simulation of Behaviour.

Maturana, H.R., Varela F.J. (1987). *Autopoiesis and cognition: The Realization of the Living*. (R.S. Cohen, M.W. Wartofsky, Eds.). D. Reidel Publishing Company, The Netherlands. https://monoskop.org/images/3/35/Maturana_Humberto_Varela_Francisco_Autopoiesis_and_Congition_The_Realization_of_the_Living.pdf

Mitchell, M. [@MelMitchell1] (2022, June 11). *Such a strange article. It's been known for *forever* that humans are predisposed to anthropomorphize even with only the shallowest* [Tweet; quotation tweet]. Twitter. https://twitter.com/MelMitchell1/status/1535628390509686785

Ney, A. (n.d.). *Reductionism*. Internet Encyclopedia of Philosophy. https://iep.utm.edu/red-ism/

Parthemore, J., Whitby, B. (2012). Moral Agency, Moral Responsibility, and Artefacts: What Existing Artefacts Fail to Achieve (and Why), and Why They, Nevertheless, Can (and Do!) Make Moral Claims Upon Us. In D. Gunkel, J. Bryson & S. Torrance (Eds.), *THE MACHINE QUESTION: AI, ETHICS AND MORAL RESPONSIBILITY,* (pp. 8–32). The Society for the Study of Artificial Intelligence and Simulation of Behaviour.

Prasad, D. (2008). Content Analysis: A method in Social Science Research. In D.K Lal Das and V. Bhaskaran (Eds.), *Research methods for Social Work,* (pp. 173-193). New Delhi. http://www.css.ac.in/download/deviprasad/content%20analysis.%20a%20method%20of%20social%20science%20research.pdf

Proyas, A. (Director). (2004). I, Robot. [Film]. 20th Century Fox.

Rosas, A. (2012). The holy will of ethical machines: a dilemma facing the project of artificial moral agents. In D. Gunkel, J. Bryson & S. Torrance (Eds.), *THE MACHINE QUESTION: AI, ETHICS AND MORAL RESPONSIBILITY,* (pp. 8–32). The Society for the Study of Artificial Intelligence and Simulation of Behaviour.

Sebeok, T. (2001). Signs: An Introduction to Semiotics (2.ed). Toronto Buffalo, London. https://circulosemiotico.files.wordpress.com/2012/10/sebeok_thomas_signs_an_introduction_to_semiocs_2nd_ed_2001.pdf

Shapiro, L., Spaulding, S. (2021). *Embodied Cognition*. The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/win2021/entries/embodied-cognition

Sputore, G. (Director). (2019). I am Mother. [Film]. Netflix.

Sung, J., Guo, L., Grinter, R., Christensen, H. (2007). *"My Roomba Is Rambo": Intimate Home Appliances.* International Conference on Ubiquitous Computing. DOI:10.1007/978-3-540-74853-3_9

Vallance, C. (2022, June 13). *Google engineer says Lamda AI system may have its own feelings.* BBC. https://www.bbc.com/news/technology-61784011

Whitby, B. (2008). *Sometimes it's hard to be a robot: A call for action on the ethics of abusing artificial agents*. Interacting with Computers. doi:10.1016/j.intcom.2008.02.002