A.I.: Artificial Intelligence as Philosophy: Machine Consciousness and Intelligence

David Gamez

Department of Computer Science, Middlesex University, London, UK
david@davidgamez.eu / www.davidgamez.eu

Abstract

A.I.: Artificial Intelligence tells the story of a robot boy who has been engineered to love his human owner. He is abandoned by his owner and pursues a tragic quest to become a real boy, so that he can be loved by her again. This chapter explores the philosophical, psychological and scientific issues that are raised by the film. It starts with *A.I.*'s representation of artificial intelligence, and then covers the consciousness of robots, which is closely linked to ethical concerns about the treatment of AIs in the film. There is a discussion about how *A.I.*'s interpretation of artificial love relates to scientific work on emotion, and the chapter also examines connections between the technology portrayed in *A.I.* and current research on robotics.

Keywords: A.I., A.I.: Artificial Intelligence, Supertoys last all summer long, Spielberg, Adliss, Kubrick, intelligence, artificial intelligence, Turing test, AI, consciousness, artificial consciousness, machine consciousness, emotion, love, imprinting, robots, uncanny valley, mecha, toy, companion, developmental robotics, Luddite

1. Introduction

A.I.: Artificial Intelligence (2001) is based a short story, *Supertoys Last All Summer Long*, by Brian Aldiss (2001). The film rights for *Supertoys Last all Summer Long* were purchased by Kubrick, who collaborated with several writers and the graphic artist Chris Baker on developing the story into a film (Baker et al. 2009). Kubrick intended to use robotics and special effects to create David, but these were not sufficiently advanced at that time, so he shelved the project. Kubrick knew Steven Spielberg and wanted him to direct *A.I.*, both because he thought that Spielberg would do a better job and because Spielberg worked faster, so the aging of a real boy cast as David would not be an issue during the shoot. Kubrick took up the project again after *Jurassic Park (1993)* demonstrated that special effects were good enough to realize his vision. When Kubrick died in 1999 the project was passed to Spielberg, who wrote the screen play and directed the final film. Most of the key ideas in the film are derived from Kubrick's development work, including the strange final section with the resurrection of Monica. The film's visual design was largely based on the drawings by Chris Baker, who worked closely with Spielberg on the project.

Kubrick had a strong interest in fairy tales and believed that they embody profound themes of human existence. The story of Pinocchio is central to Kubrick's adaptation of Aldiss' stories and Kubrick often referred to the project as his Pinocchio film. The original Pinocchio is a selfish naughty boy, who is more focused on pleasure and adventure than pleasing his father. He doesn't want to study or work and gets endlessly distracted by people who make false promises. Becoming a real boy is not important to him; it is only after he is willing to work hard and shows tenderness towards his father that he is made into a real boy by the Blue Fairy (Collodi 1995). In *A.I.* David starts out as a loving slightly cheeky boy whose rejection by Monica unleashes ugly emotions, such as the violence he exhibits towards his copy. After his abandonment David sets out on a monomaniacal quest to become a real boy so that he can recover Monica's love – a path he is tragically committed to by the imprinting process. David's resolution and fixity of purpose is very different from Pinocchio's picaresque pleasure seeking.

A.I. has good special effects, impressive visual design, some great acting and excellent cinematography. However, the plot has many flaws and the film will probably never achieve the status of a classic. Despite the title, it is not really a film about AI: it just takes human-level AI for granted and ignores the many forms and levels of artificial intelligence that are likely to be present in the future. It deals with the ethical treatment of robots in an interesting way, but this should have been linked with artificial consciousness, not artificial intelligence. The central part of the film is a well-told story about a robot abandoned by the person he has been engineered to love. Then we are transported into a distant future populated by advanced robots (easily mistaken for aliens) who facilitate a tragic 'happy' ending in which the robot's fate is left uncertain and the person he loves disappears forever. In the early scenes Haley Joel Osment does an excellent job of portraying the creepy uncanniness of the robot, David. Then the film gives up on uncanniness and the robots are scripted with ordinary human psychology.

This chapter explores the philosophical, psychological, and scientific questions asked by *A.I.* After the plot summary in Section 2, Section 3 discusses the film's representation of intelligence and examines its treatment of AI technology and AI safety concerns. Section 4 covers natural and artificial consciousness and explains why the ethical treatment of AIs should be linked to their consciousness, not to their intelligence. Emotions play a key role in *A.I.*, so Section 5 explains how psychological theories of love and imprinting connect with the portrayal of artificial emotions in the film. The final section explores *A.I.*'s vision of robot technology.

2. The Plot of *A.I.*

A.I. is set at a time when rising oceans have drowned cities and displaced millions of people. To address a lack of resources, governments have introduced strict controls on reproduction: potential parents need a license to have a child and few licenses are issued. Robot technology has made considerable progress and humanoid robots (called "mechas") are common in society. The claim is rather implausibly made that mechas consume less resources than people.

The film opens with a research meeting at the robot manufacturer Cybertronics. Professor Hobby introduces Sheila, an example of the current generation of robots. She looks human, but she does not have human emotions. When Hobby stabs her in the hand she interprets the injury as physical damage, not pain. She defines love as a series of behaviors, such as widening her eyes and quickening her breath. Like all the current mechas, she is based on "neuron sequencing technology". This generation of mechas is good at imitating human behavior, but they feel nothing inside. Hobby then outlines his plan to build a new generation of mechas that genuinely love the people that they imprint on. This is vaguely linked to "work on mapping the impulse pathways in a single neuron" - presumably some kind of advanced brain-inspired technology. The new mechas are intended to be child companions, targeted at couples waiting for a license to have a child. The scene ends with a pertinent question from one of the team: "If a robot could genuinely love a person, what responsibility does that person hold toward that mecha in return?"

The next scene is a cryogenic hospital in which Martin, the real child of Henry and Monica, is preserved in a state of suspended animation until a cure for his fatal illness can be found. Henry works at Cybertronics and, because of his special situation, he and Monica are given the first prototype of the new mecha child, David. David's initial behavior is pretty creepy: he never blinks, he has a very straight posture, he appears suddenly without warning, and he has bouts of exaggerated and inappropriate laughter. Gradually Monica gets used to him and Henry explains the imprinting process: when a sequence of random words is read to the mecha, he will love the adoptive parent. This imprinting is irreversible: if the mecha is no longer wanted, it must be sent back to Cybertronics to be destroyed. After imprinting Monica and David start to bond and Monica gives Martin's toy robot, Teddy, to David.

The positive relationship that develops between Monica and David is disrupted by Martin's recovery and return home. Martin and David compete, as boys do, and Martin uses David's artificiality against him - describing him as a "super toy". David does not appear to fully understand that he is a mecha, and Martin asks cruel questions that emphasize David's artificiality and Martin's superior status as a real boy. Martin asks Monica to read Pinocchio to them and pressures David into eating human food, which causes him to break.

Later Martin tells David that he has a special mission for him, which will make Monica love him, but David has to promise to carry out the mission before he is told what it is. David promises and Martin tells him that he has to cut off a lock of Monica's hair. He says that he is not allowed, which is presumably some kind of AI safety programming, similar to Asimov's laws of robotics (see Section 3.4). However, Martin made him promise to do it, and presumably keeping promises is part of his programming too. So David tries to cut off a piece of Monica's hair, but she wakes and David nearly pokes her in the eye with the scissors. At Martin's birthday party the other children ask David humiliating questions and test his damage avoidance system by threatening him with a knife. David gets scared and holds onto Martin. They accidentally fall into the pool and Martin nearly drowns.

The pool incident makes Monica decide that David can no longer live with them. She does not want to send David back to Cybertronics for disposal, so she abandons David in the woods with Teddy. David asks if he can come home if he becomes a real boy. Monica says that he is not real and that stories like Pinocchio are not real. In the woods David meets a sex robot, Gigolo Joe, who has been framed for murder. They are captured and taken to the Flesh Fair: an arena in which robots are destroyed in spectacular ways for the entertainment of the crowd. David and Joe are placed beneath buckets of acid and the crowd are invited to throw bean bags to release the acid. However, the crowd are unwilling to destroy them because of David's high level of realism and his apparent age. So they throw the bean bags at the Flesh Fair host and David and Joe escape.

As David and Joe wander through the woods, David explains his plan to find the Blue Fairy, who, he hopes, will change him into a real boy, so that Monica will love him. Joe suggests that they go to Rouge City to consult a search engine called Dr. Know about the location of the Blue Fairy. Dr. Know gives them an enigmatic clue that leads them to Manhattan, where they discover a large building with a laboratory and library inside. Sitting in a chair is another mecha, exactly like David. The second David is friendly, but the first David perceives him as a rival for the love of Monica. In a fit of rage the original David smashes the other mecha to pieces. Professor Hobby appears and explains that the answer from Dr. Know was planted to bring David to his lab. He describes David as special and unique and leaves the room to find the other members of his team. David wanders through the lab and sees many versions of himself in different stages of assembly as well as boxes containing copies of himself ready to be shipped. In despair he throws himself off the building into the sea, where he sees a submerged fair and a statue of the Blue Fairy. Joe rescues David, but then is taken away by the police and David and Teddy return to the submerged fair in an amphibicopter and park opposite the Blue Fairy. A falling Ferris wheel traps the amphibicopter and David is left staring at the Blue Fairy, constantly repeating his request to become a real boy.

Eventually the ocean freezes and David stops moving. Thousands of years later advanced super-mechas are carrying out archaeological excavations on the site. These super-mechas look like aliens, but the intention of the film is that they are advanced robots that have taken over after humanity died out. David is revived and meets something that looks like the Blue Fairy in a reconstruction of Monica's home. David asks the Blue Fairy to make him into a real boy. She replies that this is impossible and explains that Monica is dead but can be temporarily resurrected from her DNA. Teddy gives the Blue Fairy the lock of Monica's hair that David cut off in an earlier scene.

The resurrection of a person from their DNA only works for one day. After that the person falls asleep and cannot be resurrected again. Despite this limitation David insists that Monica is brought back and when she wakes they spend a perfect day together. The resurrected Monica loves David and David has the happiest day of his life. At the end of the day David falls asleep and goes "to that place where dreams are born." There is some ambiguity about whether this resurrection is just a dream and this is likely to have been the original intention of Kubrick (Baker et al. 2009). However, in the final film it seems reasonably clear that the living Monica is really brought back for one day and then dies forever. The final fate of David is left undetermined. Perhaps he 'dies' too, although he surely could be kept going by super-mecha technology. A more likely scenario is that he would be condemned to a lonely existence in which he loves Monica and grieves for her while struggling to adapt to the alien world of the super-mechas.

3 Intelligence

3.1 Natural and Artificial Intelligence

The mechas in *A.I.* have human-level artificial intelligence: they can perceive and identify objects, navigate in their environments, understand natural language, reason, plan, and so on. Artificial systems are usually judged to be intelligent if they exhibit behaviors that require intelligence in natural systems. Many definitions of intelligence have been put forward, including cognitive ability, rational thinking, problem-solving, goal-directed adaptive behavior and an ability to make accurate predictions (Gamez 2021; Legg and Hutter 2007a). Some of these definitions are anthropocentric; others apply to many different types of system.

Some people believe that human intelligence is completely general and can tackle *any* problem. The idea of artificial general intelligence (AGI) is usually derived from beliefs about the generality of human intelligence. However, humans do not have a completely general intelligence - for example, we cannot reason about large data sets or mentally manipulate five-dimensional objects. As Chollet (2019) points out, the human brain evolved to help us survive in a hunter-gatherer environment and it has a limited ability to generalize beyond this environment. If human intelligence is not completely general, then there is very little reason to believe that a completely general artificial intelligence is possible.

A more plausible view is that there are many different types of intelligence that are optimized for different environments. Some intelligences are good at chess; others excel at ATARI video games. This idea has often been discussed in the literature on intelligence. For example, Gardner (2006) claims that there are multiple types of intelligence, including musical intelligence, linguistic intelligence and emotional intelligence. Warwick (2000) frames this more generally with his idea that intelligence is a high-dimensional space of abilities.

Within AI research there is a popular distinction between narrow and general artificial intelligence. Systems that exhibit one type of intelligence are often called "narrow" - for example a chess playing program is a narrow AI because it cannot play draughts or Monopoly. Narrow intelligence is usually contrasted with general intelligence, but if general intelligence is a myth, then all the intelligences that we know or can imagine are, to a greater or lesser extent, narrow and we will never be able to build a completely general AI. However, we can still compare systems according to the narrowness/generality of their intelligence. Suppose we have ten environments of similar complexity. An intelligence that performs well in eight of these environments is more general (less narrow) than an intelligence that only works in one environment.

Intelligence is a *functional* property that can be implemented in many different ways. For example, a given piece of intelligence can be implemented using biological neurons, simulated neurons, computer programming, clockwork, and so on - the physical details are irrelevant as long as the system behaves in a particular way. This is very different from consciousness, which is closely tied to the physical nature of a system (see Section 4.1).

3.2 The Measurement of Intelligence

How could we measure the intelligence of the mechas in *A.I.*? Human intelligence is often measured using IQ tests, which have verbal, spatial-reasoning and mathematical questions. Verbal, spatial, and mathematical abilities are thought to be linked to intelligence, so humans that perform well on these tests are thought to be more intelligent than people who perform less well. It is often claimed that IQ tests only measure the ability of people to perform IQ tests, not intelligence itself. However, the results of IQ tests correlate with other indicators of intelligence, such academic grades, publication of scientific papers and success in professional careers (Robertson et al. 2010).

Animals cannot take human intelligence tests, so there has been a lot of work on the development of cognitive test batteries for animals (Shaw and Schmelz 2017). While it might be possible to come up with a plausible set of tests that could be applied to similar animals, this approach is likely to neglect the different types of intelligence that animals develop to survive in their ecological niches. A measure of intelligence that is designed for sheep or fish, for example, cannot easily be transferred to birds or bees. A second problem with the measurement of non-human animal intelligence is that we do not have a way of connecting an animal's test results to other indicators of intelligence for that species. Most people would agree that a person who gets top grades in school, gets a first at MIT and publishes ground-breaking physics research is likely to be intelligent. If an intelligence test gives this person a low score, then this is a failure of the test, not an indicator of low intelligence. But how could we ground the results of intelligence tests in octopi, bees or dogs? Animals do not take advanced degrees or write papers on quantum theory. It is far from clear how we could prove that intelligence tests in animals measure anything more than the ability to perform the test itself. These problems get worse when we try to use batteries of tests to measure intelligence in artificial systems. AI systems can be programmed to pass IQ tests that are designed for humans. However, IQ tests were designed to measure a more general ability in humans, whereas an AI that is programmed to get high scores in IQ tests cannot do anything except get high scores on IQ tests.

The Turing test was originally proposed as a way of answering the question of whether a machine could think. Turing (1950) described a thought experiment in which a human and a machine were connected to an electronic typing system and placed in a separate room. A human tester asked the two systems questions and tried to decide which was the human and which was the machine. If the human tester could not reliably identify the machine, then the machine would be judged to be capable of thinking. Thinking is not of much interest to modern AI researchers, so most people view the Turing test as a way of establishing whether a machine is as intelligent as a human. Many variations of the Turing test have been developed, including behavior in game environments (Hingston 2009) and the Animal-AI Olympics (Crosby et al. 2019), which provides an environment in which artificial systems can attempt tasks that are believed to require intelligence in animals. If David took the Turing test, he would pass most of the questions posed by unskilled examiners. However, some aspects of his mind could easily be identified by a skilled interrogator. For example, a real boy would have memories about earlier childhood and David is likely to be better at mathematics than a child of a similar age. Turing testing has the limitation that it can only determine whether a machine's intelligence is identical to human intelligence. It cannot measure non-human forms of intelligence or intelligence that exceeds human levels.

People have developed *universal* measures of intelligence that, in theory, can be applied to any system at all. For example, Legg and Hutter (2007b) developed a universal measure that sums the rewards that an agent receives across all possible environments, with some adjustment for the complexity of different environments. Hernández-Orallo and Dowe's (2010) algorithm is based on inductive inference, prediction, compression and randomness. Gamez's (2021) measure is based on the number of accurate predictions that a system makes in a set of environments. In the future universal measures of intelligence could become powerful tools for comparing intelligence in different types of natural and artificial system.

3.3 AI Technology

At the start of *A.I.,* Professor Hobby states that the current mechas are based on "intelligent behavioral circuits, using neuron sequencing technology." The next generation, including David, was to be based on "mapping the impulse pathways in a single neuron." Hobby appears to be suggesting that the mechas' minds are based on simulated brains that have been modified to produce specific behaviors, such as David's imprinting.

Many different approaches have been used to build intelligent machines. The earliest systems were constructed with mechanical components. For example, Jacques de Vaucanson's Digesting Duck could flap its wings, drink water, and pretend to digest grain. Clockwork has also been used for mathematical operations. The Antikythera mechanism (around 100 BCE) could predict astronomical positions and eclipses and in the 19th Century Charles Babbage's Difference Engine carried out polynomial calculations using cogs and gears. Babbage also designed a programmable mechanical computer called the Analytical Engine. His collaborator, Ada Lovelace, wrote programs for the Analytical Engine and suggested novel non-mathematical applications for it, such as music composition. The Digesting Duck, Antikythera mechanism and Analytical Engine were impressive achievements. However, the cost, speed and unreliability of clockwork limit its usefulness for building complex AI systems.

The development of electronic computers created the modern field of artificial intelligence. In the early days AI programs were sets of rules that specified actions to take when certain inputs were detected. For example, Terry Winograd's SHRDLU was a program that interacted with a virtual world, which contained blocks and cones. It could answer questions about the blocks and cones in natural language and change the world on command. These early AI systems often used a search procedure to find solutions to a problem or to plan actions. However, it was soon realized that more realistic and complicated problems had a massive search space, which could not possibly be explored. The early AIs also could not handle minor variations in their environment that had not been anticipated by the programmer. Suppose a cake recipe specifies one large egg, but there are only two small eggs in the fridge. A human would add the two small eggs to the mixture. A rule-based AI could only bake the cake if the two-egg scenario had been anticipated by the programmer.

People tried adding more rules to cope with more situations, but it rapidly became clear that high levels of intelligence could not be achieved with this method.

Computers can learn about the world and use this learnt knowledge to plan actions. Many machine learning approaches have been developed, including statistics, genetic algorithms and support vector machines. Today the most successful machine learning method is a deep neural network, which has multiple layers of simulated neurons that can be trained on millions of pieces of data. AI systems based on deep networks can reach human-level performance on classification tasks, such as face recognition, and outperform humans on games, such as Breakout and Go. The intelligence of machine learning systems is constrained by the data that they have been trained on and they often have a limited ability to work outside this context. For example, a deep network that has been trained for face recognition is incapable of processing natural language. So learning by itself does not solve the problem of the brittleness of AI systems compared to humans. In some AIs machine learning is combined with rule-based approaches – for example, a self-driving car might use deep networks for classification tasks, such as object identification, and hard-coded rules to control which actions to take under specific circumstances - if a pedestrian is in front of the vehicle (identified with deep network), turn on the brakes (AI rule).

Many AIs have been developed that use natural language processing (NLP) to understand and respond to human input. In the early days people built simple chat systems, known as chatbots, that were entirely hard coded. The chatbot had a set of input-output rules: when the input matched a rule, it responded with the corresponding output. When no match was found, the chatbot would output a question or attempt to change the subject. Some modern chatbots, such as Alexa and Siri, are based on this technology. More recently people have been building chatbots with deep neural networks that are trained on large quantities of text data from the Internet. These systems can answer questions and generate text in a more dynamic convincing way.

Many contemporary researchers are interested in the possibility that human-level AI could be built by scanning the neurons and synapses in a human brain and simulating them in a computer. This could be a way of creating an AI with human-level intelligence without the complex training and design process that goes into our current AIs. Simulated brains can work faster than biological brains and many copies can run simultaneously. Dead brains can be scanned by cutting them into very thin slices, taking pictures with an electron microscope and building a three-dimensional model of the neurons and connections from the scanned images. This process is slow, but it might eventually become possible to identify the structure of the 100 billion neurons in the human brain and their 10^{15} connections. So far scientists have managed to map the 100,000 neurons and 10^9 connections in a cubic millimeter of mouse brain. Large-scale simulations of millions of neurons have been built, but these are only very rough approximations to the human brain. More accurate simulations with tens of thousands of neurons have been created and people are developing dedicated neuromorphic hardware that will enable us to run brain simulations more efficiently. In time it is conceivable that we will be able to improve the scanning accuracy and scale up our ability to simulate the brain at a high level of detail. However, the brain is a very complex system with vast numbers of chemical and electrical feedback loops. Even if we had the neural data and the computation capacity it would take many years to get a simulated brain to work in the same way as a living biological brain.

3.4 AI Safety

At Martin's birthday party, other (real) children threaten David with a knife, and he grabs hold of Martin in fear. They both fall into the pool and Martin is nearly drowned. In another scene David tries to cut off a lock of Monica's hair and nearly pokes out her eye with the scissors. These incidents dramatize important concerns that people have raised about AI safety and liability.

There are at least four reasons why AIs could harm humans:

1. *Deliberate*. An AI kills a human because it has been programmed to kill humans. Military robots fall into this category as well as AIs that are programmed to use force to protect someone from harm – for example, a bodyguard robot.

2. *Misperception*. The AI misperceives the situation and acts according to its misperception. As far as the AI is concerned it is taking care not to hurt humans, but because it has interpreted the situation incorrectly, it ends up doing harm. For example, a butler robot might think that a baby on a table is a roast turkey and carve it up.

3. *Unresolvable dilemma*. Whatever the robot does leads to harm. This was the case with David, who appeared to have two rules: 1) Do not harm humans. 2) Keep promises. By making him promise to do an unknown act (cut off a lock of Monica's hair), David is forced to break one of his AI safety rules. In this case his desire to be loved by his mother led him to choose to break the rule about not harming humans.

4. *Accident*. An industrial robot arm hits a worker that strays into its path; a self-driving car takes a corner too fast and ploughs into some pedestrians. David's near-drowning of Martin falls into this category.

Some people believe that the dangers posed by AIs could be reduced if we could hard-code safety rules into them. The most famous set of safety rules that have been put forward are Asimov's (1952) laws of robotics:

1. Robots shall not harm a human, or by inaction allow a human to come to harm.
2. Robots shall obey any instruction given to them by humans.
3. Robots shall avoid actions or situations that could cause them to harm themselves.

While these laws initially seem plausible, they cannot protect people against misperception, unresolvable dilemmas, and accidents. Asimov was well aware of this, and *I, Robot* explores the many ways in which these apparently simple laws fail to produce desired behaviors.

AI safety is particularly challenging in machine learning systems because it can be hard to understand what they have learnt and difficult to predict how they will behave in new or unexpected situations. This is often an issue with deep neural networks that have been trained on millions of pieces of data. An AI that cannot be understood by humans is referred to as a black box. There is ongoing work to try to white box AI systems, so that we can understand what is going on in their "minds" and address potential safety issues.

There are also complex questions about AI liability, particularly with machine-learning systems. If a self-driving car fails, is it the owner's fault or the manufacturer's fault? When David violently destroys a copy of himself, is this because of his programming or because he has been mistreated by his environment? White-boxing AIs could make it easier to identify the causes of unwanted behavior. In the future robot manufacturers might be required to build systems that behave well (Russell 2019). However, this is not at all easy because human morality is ambiguous and people often have incompatible sets of values (Haidt 2013). We need to solve our own moral relativism before we can conceivably program morality into a robot.

4. Consciousness

4.1 Natural Consciousness

The only significant mention of consciousness in *A.I.* is when Hobby claims that "love will be the key by which they acquire a kind of subconscious never before achieved. An inner world of metaphor, of intuition, of self-motivated reasoning. Of dreams." Hobby is not suggesting that he will build a conscious robot, who experiences the colors, sounds, and smells that are typically thought to constitute conscious experience. Instead, a more liminal world of imagination and dreams will be given to a robot that may or may not be fully conscious.

Consciousness is often defined as the stream of colorful noisy smelly tasty experiences that starts when we wake up in the morning and disappears when we fall into dreamless sleep at night. The modern concept of consciousness emerged in Europe in the 17th Century (Wilkes 1988). Prior to this, people believed that conscious experiences were objective properties of the physical world – green was attributed to trees, not to the interaction between light, trees, eyes and brain (a position known as naïve realism). The renaissance of atomism in the 17th Century led to the physical world being interpreted as a realm of colorless atoms bouncing about in the void. When these atoms interacted with our senses they produced conscious experiences of colors, smells, etc. This led to a distinction between primary qualities, which were properties of the atoms (for instance, size, shape and speed), and secondary qualities, which were properties of consciousness (for example, color, smell and sound). Primary qualities were thought to be physically real, but 17th Century thinkers could not ignore their experiences of the colorful smelly world that they lived in from day to day. So the concept of consciousness was developed to accommodate people's experiences of colors, smells and sounds, which had been squeezed out of the physical world by atomism (Gamez 2018).

Since the 17th Century we have been attempting to put consciousness and the physical world back together. Some people have tried to reduce consciousness to the physical world (a position known as physicalism), but it makes no sense to claim that colorful experiences are identical to neuron activity. Other people have taken consciousness to be the primary reality – for example, idealists, like Berkeley (1957), or phenomenologists like Husserl (1960), who suspended belief in the physical world. A better way out of this dilemma is to accept that consciousness and the physical world are both real and to scientifically study the relationship between them. This type of research is carried out by people searching for the neural correlates of consciousness. These scientists measure the brain, measure consciousness and look for neural activity patterns that only occur when the brain is conscious (Koch et al. 2016). There are also quantum and electromagnetic theories of consciousness, so work on the neural correlates of consciousness can be generalized into a search for physical patterns that are correlated with the presence of consciousness and, ideally, with specific conscious contents.

Many people have suggested that consciousness is correlated with computational or informational patterns (Cleeremans 2005; Tononi 2008), but there are strong reasons for thinking that computation and information are subjective interpretations of a physical system (Gamez 2018). Whether or not something is conscious cannot depend on how we interpret it, so it is much more likely that consciousness is correlated with specific physical patterns – for example, patterns in biological neurons, electromagnetic waves or quantum states.

As Popper (2002) points out, a final theory of the relationship between consciousness and the physical world will not be a long list of correlations. Ideally, we would like to find a compact mathematical theory that maps between descriptions of the physical world and descriptions of conscious states. Such a theory could generate a description of the conscious state that is associated with a particular physical state. Or, conversely, if we knew the conscious state, then it should be able to generate a description of the corresponding physical state (Gamez 2018). This is illustrated in Figure 1.

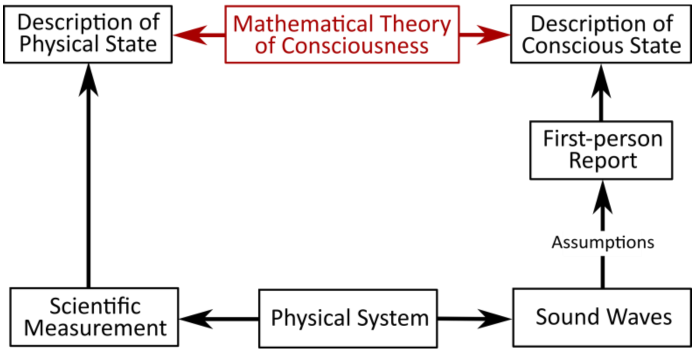
Description of Physical State
Mathematical Theory of Consciousness
Description of Conscious State
First-person Report
Assumptions
Scientific Measurement
Physical System
Sound Waves

Figure 1. *Mathematical theory of consciousness*. Scientific measuring instruments (fMRI scanners, EEG, etc.) are used to measure the system and produce a description of its physical state. The system also produces sound waves (and other behavior) that we can interpret as first-person reports about consciousness by making certain assumptions (Gamez 2018). These first-person reports are then converted into a description of the system's conscious state. A mathematical theory of consciousness describes the relationship between physical and conscious states. It can generate a description of consciousness from a description of a physical state and generate a description of a physical state from a description of consciousness.

Progress is being made with the development of mathematical theories of consciousness. The most popular is Tononi's information integration theory of consciousness, which converts information patterns into a prediction about the amount of consciousness, the location of the consciousness and the structure of the consciousness (Oizumi et al. 2014). There is little experimental support for Tononi's algorithm and it has severe conceptual and performance issues (Gamez 2016). However, it does illustrate the form that a mathematical theory of consciousness could take.

4.2 Artificial Consciousness

Although consciousness is barely mentioned in *A.I.*, it is one of the most important aspects of the plot. Suppose that David was a purely mechanical system with no trace of consciousness. He has the external appearance of a sad boy who wants his mother to love him. He makes certain sounds – for example, when he is abandoned, he pitifully pleads "Why do you want to leave me? Why do you want to leave me? I'm sorry I'm not real, if you let me I'll be so real for you!" But if David is *not* conscious, then these are just sounds produced by mechanical processes, no different from the vibrations in the air produced by Vaucanson's flute player. Our belief that David consciously experiences love, fear and pain drives our empathy for him in his adventures. In the Flesh Fair we respond to his imminent destruction with far more emotion than we would if we were watching a washing machine being destroyed.

Artificial consciousness is a complex field that can be divided into four different areas (Gamez 2018):

1. *Replication of external behaviors that humans exhibit when they are conscious*. For example, humans only respond to novel situations and execute delayed reactions to stimuli when they are conscious. These behaviors can be exhibited by AI systems.
2. *Models of the correlates of consciousness*. For example, researchers have created simulations of the neural correlates of consciousness (Shanahan 2008).

3. *Models of consciousness*. The structures of consciousness have been documented by phenomenologists like Husserl (1960). These can be modelled in computers and used to control robots (Gravato Marques and Holland 2009).
4. *Artificial systems with something that corresponds to our conscious experiences*. These Ais would have something like the colors, tastes sounds and smells that appear when we wake up in the morning and disappear when we fall into deep sleep at night.

In *A.I.* David replicates the external behaviors of a conscious human. His mind might be based on a model of the neural correlates of consciousness or on a model of consciousness. From the point of view of David's treatment in the plot, the most important question is whether he actually has conscious experiences – something similar to the colors and sounds that we experience when we are conscious.

In humans it is straightforward to infer consciousness from external behavior. When my eyes are open and I am speaking coherently, people naturally conclude that I am having colorful smelly noisy conscious experiences. With occasional exceptions, such as epileptic automatism, the inference from human behavior to consciousness is judged to be reliable and it is the basis for medical diagnoses of consciousness, such as the Glasgow Coma Scale (Teasdale and Jennett 1974). This inference works in humans because we assume that most people are built in the same way: if they are behaving in the way that we do when we are conscious, then it is reasonable to assume that they are conscious too.

External behavior is not a correlate of consciousness in humans. It is the neural patterns that lead to this external behavior that are correlated with consciousness. In a robot there are an infinite number of different ways of producing a given piece of behavior. For example, the phrase "I am conscious" could be output by a single line of code, a biological brain or by a sophisticated chatbot trained on hundreds of millions of pieces of data. Some of these systems might be conscious; many are highly unlikely to be conscious. Judgements about an artificial system's experiential consciousness cannot be based on its external behavior.

In *A.I.* David behaves like a human. If he was human on the inside, we would have no hesitation in attributing consciousness to him. But David is a robot and the consciousness of the mechanisms that produce his external behavior cannot be inferred from that behavior. If we want to know whether David is *really* conscious, we have to use a theory of consciousness that can reliably map between physical and conscious states (see Figure 1). We can use this mathematical theory to convert a description of David's physical states into a description of his conscious states, as shown in Figure 2.

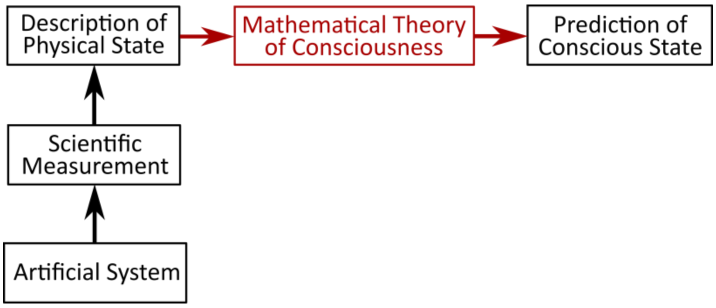
Description of Physical State
Mathematical Theory of Consciousness
Prediction of Conscious State
Scientific Measurement
Artificial System

Figure 2. *Prediction of artificial system's conscious state*. Scientific instruments measure the physical state of the artificial system. A reliable mathematical theory of consciousness converts the physical description into a prediction of the artificial system's consciousness.

This approach has limitations because a theory of consciousness that has been developed on natural systems might miss relationships between physical and conscious states that do not exist in natural systems. However, it is the *only* reliable way of making inferences about the consciousness of artificial systems. *A.I.* encourages us to relate to David in the same way that we would relate to a conscious human boy. However, if David was a real robot, it would be a completely open question whether he was conscious.

4.3 Simulation, Intelligence and Consciousness

People often confuse computer simulations with the thing that is being simulated, particularly when it comes to robots and AI. Suppose we want to model lynx and rabbit populations. Lynxes eat rabbits and both animals reproduce at a finite rate with influences from resources, diseases, and so on. These interactions can be captured by a simple computer program. When the program runs, the lynx and rabbit numbers are stored as voltage patterns in the memory of the computer. In practice these voltage patterns are constantly fluctuating, but for simplicity we will treat them as 1s and 0s. So if there were 56 rabbits, 111000 would be in the computer's memory. The connection between 111000 and real rabbits is made in our minds: 111000 is not warm or furry, it does not eat grass and it cannot be eaten by a lynx.

Voltage patterns *can* produce intelligent behavior. David's intelligence is a *functional* property: anything that behaves intelligently is intelligent. Intelligent systems can be implemented with voltage patterns, people manipulating Chinese characters in a room (Searle 1980) or by ant colonies (Hofstadter 1979). Any of these systems could, in theory, control a robot like David that is as intelligent as a human being. There is no distinction between a simulation of intelligence and intelligence itself.

Consciousness is a *physical* property of a system (see Section 4.1). The way in which a system is physically built matters for consciousness. A computer simulating a brain and a biological brain might produce the same behaviors in a humanoid robot or biological body. This does not mean that they have the same consciousness. Only a reliable mathematical theory of consciousness can decide whether non-biological systems are conscious. A mecha controlled by simulated neurons is very unlikely to be conscious in the same way as a biological human brain.

4.4 Consciousness and Ethics

Throughout the film David is treated as a thing that can be damaged with impunity. He will be destroyed by Cybertronics if his owners no longer want him. After he is abandoned by Monica, David ends up in the Flesh Fair, where unregistered mechas are wrecked for entertainment. It does not matter how David feels about being dissolved in acid – it is just the destruction of a machine that simulates feelings; the legal destruction of property that does not belong to anyone. A similar theme appears in Pinocchio when Fire-eater decides to burn Pinocchio to cook his mutton. When Pinocchio begs for mercy, Fire-eater suggests that one of the other puppets should be burnt in his place. Pinocchio is not a real boy, so it is fine to throw him into the fire to cook some meat.

At the start of the film, one of Hobby's team asks: "If a robot could genuinely love a person, what responsibility does that person hold toward that mecha in return?" Kubrick raised a similar issue: "One of the most fascinating questions that arise in envisioning computers more intelligent than men is at what point machine intelligence deserves the same consideration as biological intelligence…. You could be tempted to ask yourself in what way is machine intelligence any less sacrosanct than biological intelligence, and it might be difficult to arrive at an answer flattering to biological intelligence" (Baker et al. 2009). These ethical questions are incorrect because intelligence is a f13unctionnal property that is irrelevant to a system's treatment. As far as his intelligence goes, David's human mimicry is just a clever trick, like a special effect in the movies.

There is a close relationship between consciousness and ethics. We cannot kill fully conscious humans or cause them to suffer without their consent. The respect that we have for animals loosely correlates with their perceived consciousness: primates are given the best treatment, fish receive some consideration and insects have no protection at all. We can kill human fetuses, which are probably not conscious, but not human infants. Coma patients that are unlikely to regain consciousness are allowed to die. In most countries it is illegal to assist a fully conscious person with terminal illness to voluntarily end their life. So it is only when robot toys and companions become conscious that we are obliged to treat them in the same way as other conscious systems

When the mechas in the Flesh Fair express fear, we automatically empathize with them and attribute consciousness to them, just as we would if a real human was being destroyed. However, it is possible that none of the mechas are conscious. They could be simulating conscious human behavior with nothing going on inside. To determine whether the mechas are really conscious, we have to measure their internal physical states and use a reliable mathematical theory to convert this into a description of their consciousness. If there is no consciousness, then the mechas' destruction is merely a thrilling illusion in which something that looks and acts like a human is destroyed. On the other hand, if the mechas really are conscious, then their destruction is similar to the killing of people or animals for public entertainment.

In the future, if we want to avoid ethical issues with robot toys and companions, we could use mathematical theories of consciousness to design robots that are *not* conscious. We could then treat them like toasters and junk them when they are no longer required.

5. Emotions

5.1 What are Emotions?

The first scene of the film introduces Sheila, one of the current generation of mechas. Hobby describes her as "A sensory toy, with intelligent behavioral circuits." He injures Sheila and asks her how this made her feel. She replies that the injury was to her hand, not to her feelings. She describes love as "widening my eyes a little bit and quickening my breathing a little and warming my skin and touching …". This scene shows that the current mechas lack human emotions. They do not feel pain or love; they are just sensory toys with intelligent behavioral circuits. Hobby then suggests that Cybertronics should build a child mecha who can love: "A robot child who will genuinely love the parent or parents it imprints on, with a love that will never end." This is the axis on which the plot turns. If David merely simulated love, in the way that sex mechas simulate arousal, then he would not need to be sent back to Cybertronics for disposal. However, his genuine non-transferrable love for Monica forces him on a tragic quest to become real, so that he can be reunited with her.

Emotions are a controversial topic and many theories have been put forward. Some people believe that they are socially constructed, others think that they are natural kinds, and there are extensive debates about which feelings truly constitute emotions and which do not. In my view, the most plausible interpretation of emotions is that they are perceptions of real or virtual body states (Damasio 1994; James 2000). For example, when I see something frightening, my brain causes my skin to sweat and my stomach to tighten. My perception of these physical changes is the emotion of fear.

In humans there is not a simple one-to-one mapping between body states and emotions (Barrett 2018). The interpretation of a body state as an emotion is a complex process that partly depends on context. For example, Barrett describes how she interpreted the onset of flu as attraction during a date. Basic emotions, such as fear and joy, are probably shared across cultures; more complex emotions are likely to vary between cultures and languages, just as the partitioning of the physical world varies between cultures and languages (Gamez 2007). In non-human animals there is likely to be a more straightforward relationship between changes in body states and perceived emotions.

We learn associations between states of our environment and emotional states. Certain foods and social situations make us happy; other foods and social situations make us unhappy. Damasio describes these learnt associations as somatic markers. Somatic markers vary widely between individuals, which explains the different ways in which people seek happiness, sexual satisfaction, and so on. Some people find staring at a crucifix to be deeply fulfilling; other people experience sadness and distress at the sight of a dead man nailed to a cross.

In psychology a distinction is often made between affect and emotion. Affect is the general background feeling that you experience throughout each day: whether you are calm, bored, tired, energetic, etc. This background feeling has two components: valence and arousal. Valence is how pleasant or unpleasant you feel – for example, you might have an unpleasant stomachache or experience pleasure at the sight of your child. Arousal is how calm or agitated you are – for example, you might feel tense before a parachute jump or fatigued after a long run. There is no clear dividing line between affect and emotion. In my view the combination of arousal and valence that we experience throughout each day is more plausibly described as an ongoing emotional state that is influenced by our mind and environment. We might, for example, go on holiday to positively influence our arousal and valence in exactly the same way that we buy a new pair of shoes to make us happy.

Rationality is often celebrated as our most impressive human quality and contrasted with irrational emotions that we are supposed to master. This greatly underestimates the critical role that emotions play in our thinking. Our decisions are guided by complex interactions between real and imagined states of the world and learnt associations between these world states and the emotions that they produce in us. I go to the cinema because I have learnt an association between watching films and positive emotion states. When I think about the cinema, my brain produces a weak version of these positive emotion states in my body, which motivates me to make a plan to go to the cinema.

Emotions can be conscious or unconscious. For example, I can be stressed without consciously perceiving the signs of stress in my body. Later, when I meditate, I become aware of my hunched shoulders and the knot in my stomach. Unconscious emotions can influence actions in the same way as other unconscious perceptions. When we are considering how to treat someone, only consciously experienced emotions matter. It is wrong to cause conscious pain and fear; unconsciously perceived pain during surgery does not raise ethical issues.

5.2 Love

David's love for Monica is central to the plot of *A.I.* In humans love is a combination of emotions that can occur independently in other contexts. For example, in the early stages of a romantic relationship people often experience sexual desire, excitement, happiness and the absence of negative emotions. Love can change over time, shifting from intense euphoric states in the early stages to long term attachment. There is some debate about whether love is a basic emotion or even if it is an emotion at all (Lamy 2015).

Studies of human love often distinguish between passionate romantic love and compassionate or family love. Romantic love often includes intense states of euphoria usually with sexual desire. Compassionate love, such as the love of a mother for her child, typically has less extreme euphoria and lacks the sexual component. Both forms of love often include a suppression of negative emotions and a reduction in critical judgement (Zeki 2007).

When David is with Monica he presumably experiences warm positive feelings, pleasure in her company and an inhibition of anxiety and critical judgement. When they are apart his positive feelings decrease and his anxiety increases. Her abandonment of him presumably made the negative components more intense, strongly motivating him to be with her, so that he could experience the warm positive feelings and the inhibition of anxiety again.

5.3 Imprinting

When David arrives at Monica's and Henry's home he has a neutral attitude towards Monica and does not call her "Mommy". The imprinting consisted of reading seven words to David, which caused him to "genuinely love the parent or parents … with a love that will never end." The genuineness of David's love for Monica appears to be his unique selling point – in contradistinction to older mechas that simulate emotions without actually feeling them. Knowing that your robot 'child' genuinely and irreversibly loves you could make you feel special and increase your bonding with the robot. However, imprinting has disadvantages, which drive most of the film's plot. After imprinting David genuinely and irreversibly loves Monica, but she abandons him, and so he is driven to seek the Blue Fairy, so that he can be changed into a real boy and be loved by Monica again.

Imprinting was first scientifically described by Lorenz (1937), who observed that newly hatched geese follow the first moving object that they see. Usually this object is their mother, but they can imprint on humans or other moving objects that they are exposed to in a short period after hatching, such as a green box or football. Imprinting has been observed in birds, insects, fish and some mammals, including sheep, goats, deer, buffalo and guinea pigs (Hess 1958). Under normal circumstances, when the imprinting period is over the young can no longer form this kind of attachment or change the object or animal that they have imprinted on.

Imprinting can, in theory, be implemented in different ways. For example, the young animal could experience fear or anxiety when it is not looking at the imprinted object, or there could be a simple hardwired reflex that causes it to move in the direction of the imprinted object. The most likely explanation is that the imprinted animal starts to love the imprinted object, experiencing positive feelings and a reduction in negative feelings when it is in its presence. There is some evidence for this in sheep, where the imprinting window can be re-opened by the injection of the hormone oxytocin, which is linked to the implementation of love in the brain. In the film David's imprinting on Monica is based on love.

Imprinting could be a positive feature for robot companions. For example, we would not want to buy an expensive robot wife or husband who falls in love with someone else. Imprinting could ensure that robot companions do not betray us, abandon us or hurt our feelings. In the film the imprinting is portrayed as irreversible – David must be returned to the manufacturer for destruction if the imprinted owner no longer wants him. This has the cynical benefit for the manufacturer that they prevent a second-hand market for mechas like David, but it does have the consequence that many owners would choose to abandon their unwanted mechas to a hopeless unrequited existence, rather than send them back for disposal.

5.4 AI, Robots, and Emotions

Emotion detection helps robots to interact with humans: David wants to know if Monica is happy or sad; Gigolo Joe monitors his clients' pleasure. Emotion recognition is a well-established field in AI and there are many commercial solutions for the identification of emotions in video, voice, and text. Emotion detection is not completely reliable because there is not a simple mapping between facial expressions and emotions, and algorithms are often trained on pictures of actors who are pretending to experience different emotions.

Human-robot interaction is also easier if robots use human expressions to communicate. By simulating pleasure, Joe increases the pleasure of his clients. David's loving expressions increase his attractiveness as an artificial child substitute. Or consider a robot that is trying to teach the times tables to a child. The robot asks, "What is seven times five," the child replies, "Thirty five," and the robot replies "Correct" in a bland mechanical voice. Children will rapidly get bored with this teaching method. A robot that expresses sorrow when the answer is wrong and joy when it is correct will connect more directly with the child's emotions and engage them in learning for much longer. Early work in this area was carried out with the Kismet robot, which could perceive and express human emotions (Breazeal 2002). Many other robots that express human emotions have been developed, including the highly realistic humanoid robots discussed in Section 6.1.

In the Flesh Fair Lord Johnson-Johnson makes a strong contrast between a mecha's simulation of emotion and real emotions experienced by humans: "Do not be fooled by the artistry of this creation. No doubt there was talent in the crafting of this simulator. Yet with the very first strike, you will see the big lie come apart before your very eyes!" The 'lie' is that a mechanical system is producing the external signs of emotion to make us think that it really has emotion, whereas in fact it is just clever robotics producing an illusion. This aspect of the film is out of touch with reality because scientists have been building AIs with emotions for a long time. The simpler implementations use variables to represent emotions. The values of these variables are changed in response to input, and they are used to select behavior. This type of emotion model has been used with the AIBO robotic dog and it is part of the LIDA cognitive architecture. In more complete implementations of emotions, the robot perceives states of its real or virtual body, which change in response to external events. For example, the robot might have a stomach that changes state when a fearful stimulus is encountered. A full implementation of emotions would also enable the robot to learn new associations between body states (somatic markers) and states of the environment. In the future, AI systems with emotions are likely to become increasingly important as the critical role that emotion plays in cognition comes to be more widely recognized (Pessoa 2019).

The artificial emotions that have been implemented so far are unlikely to be associated with consciousness. For a robot to consciously experience human love it would need to generate the same bodily responses as humans in love (positive feelings, suppression of negative critical thinking, physiological responses, such as flushing and increased heart rate). This emotional response would have to be implemented in a way that is correlated with consciousness (see Section 4.2). More progress with the implementation of emotions in robots and in our scientific understanding of consciousness is required before we will be able to build robot children like David, who genuinely and consciously love the people that they imprint on.

In the film most of the current generation of mechas are far less cold and logical than Shelia. This is particularly apparent at the Flesh Fair, where we are presented with robots who do not regard their imminent destruction as something that just happens to their bodies. One mecha asks for his pain receivers to be shut down (a big contrast with Sheila at the start of the film), Gigolo Joe appears to have the full range of human feelings, and even Teddy appears to experience fear and attachment to David. This inconsistency is not surprising because the film wants us to empathize with the mechas. It would be a very different Flesh Fair if all the mechas had the same equanimity as Sheila about physical damage and calmly accepted their destruction in the same way as a microwave oven.

6. Robotics

6.1 Humanoid Robot Technology

A.I. simulates humanoid robot technology with human actors and a considerable amount of work went into the development of the robotic Teddy puppets that were used to shoot the film. In the real world, humanoid robots are difficult to control, and scientists are only just starting to build humanoid robots that can cope with realistic environments. Much more work is being done on machines that save labor in specific situations (for instance, manufacturing robots and self-driving cars), rather than on humanoid robots that directly replace human labor.

Most of the humanoid robots that have actually been built, such as ASIMO, NAO, and iCub, are rigid structures with rotating joints. They can be programmed to perform impressive feats – for example, ASIMO dances and plays football. But these behaviors are mostly pre-programmed moves within predictable environments. If the environments change by a small amount, then the behaviors often fail with unpredictable consequences. The most notable exception to this is the Atlas robot developed by Boston Dynamics, which is much better at handling variations in terrain and recovering from falls and external impacts. While most humanoid robots only mimic the external form of the human body, some also copy our musculoskeletal system, in the hope that this could lead to more natural movements and better integration between the robot's mind and body. For example, the CRONOS and ECCE robots are based on a copy of the human skeleton, which is moved by muscles modelled by cords and electric motors. These robots are much more flexible and dynamic than traditional humanoid robots and have the potential to behave in more humanlike ways. However, they are extremely difficult to control.

Work is also being done on robots that closely mimic the external appearance of humans. For example, Hiroshi Ishiguro's Geminoid robots are fairly accurate copies of particular people. There is also Ameca, developed by Engineered Arts, Sophia, developed by Hanson Robotics, and Ali-Da, which is touted as the world's first ultra-realistic humanoid robot artist. These robots look fairly convincing, they can mimic human facial expressions and they can be connected to AI technologies, such as chat, face recognition, and so on. However, none of them can move their bodies as well as Atlas: their main achievement is facial mimicry, which falls far short of the mimicry displayed by Sheila at the start of *A.I.*

Humanoid robots consume a lot of power. Research is being done to make some behaviors, such as walking, more energy efficient by copying the mechanics of the human body. However, a major breakthrough in power generation, such as the arc reactor in *Iron Man (2008)*, would be required to create the mechas in *A.I.*, which can apparently last for years without refueling or recharging.

6.2 Robot Pets, Toys, and Companions

David is a robot companion that is designed to look and act like a child and fill the emotional needs of adults who are prevented from having children. For Henry and Monica, David also fills the hole left by the illness and cryogenic suspension of Martin.

Many robot pets, toys and companions have been developed, often to provide emotional support to the elderly or hospital patients. One of the first virtual pets was the Tamagotchi, a small egg-shaped device with a screen displaying a pet that the owner could feed, train, and play games with. A nice example of a robot toy is a bear called "Super Toy Teddy", which was based on Teddy and had rudimentary AI functionality. A substantial amount of work has also been carried out on robot pets and companions for the elderly and hospital patients using a variety of robots, including NAO (small humanoid robot), Paro (robot seal) and ElliQ (robot with screen face), which was distributed to elderly people in New York.

Toys and companions are designed to manipulate our emotions. The cute visual appearance and furry squishiness of a stuffed toy induces feelings of warmth and affection. Children cuddle, love and take care of their teddies in the same way that they cuddle, love and take care of real animals. In *A.I.* David is a companion robot for adults – a much more advanced version of Paro and ElliQ – that is designed to fill an emotional hole in the lives of adults who have not been able to have children of their own.

Most toys are inert *objects* (plastic soldiers; cuddly teddies; china dolls) that can be flexibly incorporated into different play scenarios. They might break, but they never get bored; they don't care if they wear a tutu or a pirate hat. AI changes toys from objects into subjects. They have their own desires and autonomous behavior and are less adaptable for imaginative scenarios - Teddy doesn't want to play tea parties anymore because he is bored of that game. Toys that are subjects are more effective companions. A plastic dog is only intermittently constructed as a subject by its owner - when, for example, it is described as needing to eat or pee. A real dog has a strong presence and personality; its needs and desires must be taken into account. It might be willing to be dressed up and participate in games, but to a much more limited extent than a plastic dog. In *A.I.,* Teddy is a subject, a companion who operates alongside David, shares his troubles and helps him out of danger. Future AI companions could guide children's learning, act as safe confidants, and open up new possibilities for play.

Today's toys and robot companions can be treated in any way we like. We can dissect them, blow them up with fireworks or dissolve them in acid for our amusement. The film does not provide any reason for believing that David or Teddy are conscious, but it does strongly encourage us to infer consciousness in them from their external behavior. If we could use a mathematical theory to prove that they were not likely to be conscious, then it would be completely fine for them to be destroyed in the Flesh Fair for the entertainment of the crowd. However, David and Teddy are designed to engage our emotions, so it would still be upsetting to see them dissolved in acid or fired into a propeller, even if no actual harm was being done.

6.3 Developmental Robotics

In Aldiss' (2001) stories, Monica gets tired of David and Teddy because they are so predictable. David is a poor substitute for a child because he learns little and never really changes. Teddy is equally boring, playing the same games with David all summer long. If we want our robot companions to continue to engage our interest and entertain us, then they will have to learn and grow alongside their owners – an area known as developmental robotics.

Developmental robotics is an important research area because it is impossible to program AIs with human-level intelligence from scratch (see Section 3.3). While machine learning has made rapid advances in recent years, most of its successes have been in areas where large quantities of labelled training data are available, or in simple environments, such as board or video games, that can easily be modelled in a computer. Much less progress has been made with the development of AI systems that can function effectively in real-world environments – self-driving cars are probably the best example, and they have a very limited set of objectives and behaviors.

The human brain is roughly wired up at birth and it takes many years of learning to develop the intelligence of an adult human. Some researchers believe that the limitations of our current AI can be overcome by building robots that start out in an infantile state and then learn like children by being immersed in our physical and social environment. An early pioneer in this area was Grand (2004), whose robot orangutan, Lucy, could learn from her experiences. Other examples of infant and child robots are CB2, Infanoid, and Pneuborn (see Cangelosi and Schlesinger (2015) for an overview of this work). This research is at an early stage and has the problem that each training/test cycle can take a long time.

6.4 The Uncanny Valley

As robots become more humanlike, we increasingly empathize with them. Industrial robot arms do not evoke warm feelings; we can easily relate to the cute NAO robot. When robots become very similar to humans, without completely matching their appearance and behavior, then we experience unease and revulsion. Our empathy returns when robots become indistinguishable from humans. In computer science this phenomenon is known as the uncanny valley (Mori 2012).

Mori's theory of the uncanny valley nicely accounts for the levels of empathy that we experience for the characters in *A.I.* Teddy clearly looks like a toy, so we do not experience him as creepy and empathize with him. Gigolo Joe looks human and (apart from his music-playing ability) generally behaves in a similar way to humans, so we relate to him with roughly similar levels of empathy to Teddy. Monica is 100% human, and we empathize with her in the same way that we empathize with other humans. In the early stages of the film David looks human, but his behavior (in contrast to Joe) is distinctly odd – no blinking or sleeping, pretend eating, inappropriate laughter, etc. This produces a feeling of revulsion or negative empathy in us. Later in the film David evokes similar levels of empathy to Joe. This is illustrated in Figure 3.

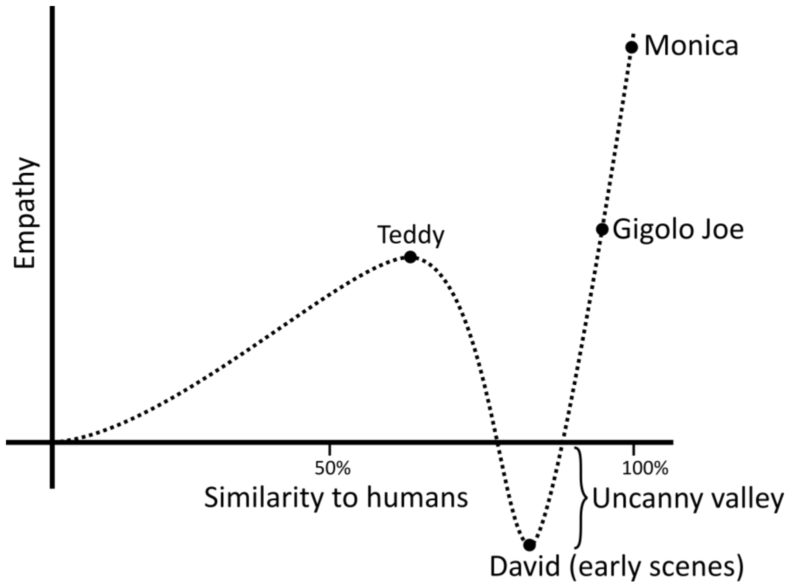

Figure 3. *The uncanny valley*. We experience negative empathy for characters that resemble humans without completely matching human appearance and behavior. More empathy is experienced for characters that are clearly artificial or that are virtually indistinguishable from humans.

There is a limited amount of empirical evidence for the uncanny valley (Kätsyri et al. 2015) and some people have applied Mori's recommendation that that we aim for the first peak initially, building systems that evoke affinity, but not so close to human likeness that we experience revulsion. We do not want to share our world with robots that creep us out, so future humanoid robots will either have to make their artificiality explicit or develop to the point at which they are indistinguishable from humans.

6.5 Real and Artificial Boys

Monica rejects David because he is a mecha who accidentally endangers her real son. But David is desperate to be with Monica and believes that she would love him if he became a real boy like Martin. So he sets off on a quest to find the Blue Fairy, who he hopes will make him into a real boy. The other mechas in *A.I.* would also benefit from becoming real. They are built to serve humans and can be tortured and destroyed with impunity when they are no longer useful. If mechas could become real (biological) humans, then they would cease to be slaves and enjoy the same rights as humans.

There is no clear dividing line between real and artificial boys. At one end of the continuum are natural womb-born children; at the other end are completely artificial systems, such as the Atlas or iCub robots, that are entirely assembled from manufactured parts. In theory David could become a real boy by replacing his artificial parts with biological parts. His mind could be built from biological neurons; a biological body could be grown for him in the lab (possibly using a combination of Monica's and Henry's DNA). The only thing that would be missing is the biological history of real boys – he would always be a mecha assembled from biological components. So even if David became a biological boy, Monica would probably never love him because he lacked the history of her natural son Martin.

Real boys have many disadvantages. They grow old and die; they have medical conditions that cannot be fixed; they cannot be backed up and restored; they are incapable of long-distance space travel. Pinocchio is much tougher than a real boy and can easily be repaired. David's mind can be backed up and his body parts can be swapped out and fixed. The super-mechas at the end of the film have outlived humanity. So David's quest to become real reverses the more common science fiction trend in which people become more artificial to enhance themselves or cheat death – *RoboCop (1987)* is a nice example. In many ways David was lucky to be an artificial boy. If he could have accepted an artificial copy of Monica, then they could have lived happily together, potentially forever.

6.6 The Displacement of Humans by Robots

The Flesh Fair is a reaction against mechas, who are accused of robbing humans of their specialness and dignity. It is a protest against artificiality and simulation: the mechas' destruction makes the lie of their appearance come apart before the crowd's eyes. The film also expresses concerns that humans are outnumbered and mechas might eventually take over. This point is nicely made by Gigolo Joe: "They made us too smart, too quick, and too many. We are suffering for the mistakes they made because when the end comes, all that will be left is us. That's why they hate us…" This prediction comes true at the end of the film when humanity wipes itself out and only the super-mechas remain. In our own society labor shortfalls are typically addressed by recruiting foreign workers and boosting the birth rate. So it is a little strange that *A.I.* presents a future in which the birth rate is cut and large numbers of power- and resource-hungry robots are built to work for humans. It is these policies, not the mechas themselves, that create a death sentence for humanity.

For the foreseeable future, humanoid robots will be too stupid and expensive to cause significant job losses. Jobs will be lost to other types of AI: self-driving cars and trucks will take drivers' jobs, call center staff will be replaced by chatbots, robots will play an increasing role in farming, and factories will become fully automated. In the past, the jobs that were lost to automation were replaced by new jobs that were created by industrialization. In the 21st Century there are few agriculture laborers and many office workers. So far the AI technology revolution is following the same pattern, but this may change as AI becomes increasingly sophisticated. If AI and robotics does result in significant net job losses, then it might become necessary to introduce some form of universal basic income – a regular payment to every member of society that covers their living costs. This could be funded by taxes on the AI and robotics companies that were responsible for the reduction in human jobs. Without a universal basic income, large numbers of unemployed people might not have enough to live on and there could be widespread social unrest.

7. Conclusions

A.I.: Artificial Intelligence is a flawed film that asks interesting questions about intelligence, consciousness, artificial emotions, and robotics. Despite the title, it is not a film *about* artificial intelligence because it takes human-level artificial intelligence for granted and does not explore the variety of forms that AI can take or the difficulties with building AI systems. The film's suggestion that the mechas' intelligence could be implemented using brain-based neural simulations is plausible and fits in with current research on biologically-inspired neural networks. AI safety issues are nicely dramatized in the early scenes and the treatment of the mechas in the Flesh Fair raises important points about the ethical treatment of AIs. However, the ethical issues raised by the film should have been linked to the mechas' consciousness, not to their intelligence.

Consciousness is almost completely ignored in *A.I.*, despite its relevance to the mechas' treatment. Humans typically infer consciousness from external behavior: something that behaves and looks like a human is judged to have human consciousness. This works with natural systems, but there are many ways in which robots can be controlled, and most of these are unlikely to be associated with consciousness. So the experiential consciousness of artificial systems cannot be inferred from their external behavior. Instead, we have to use a mathematical theory of consciousness to convert a description of the machine's physical state into a description of its conscious state. Research on mathematical theories of consciousness is still at an early stage and it will be a long time before we can make accurate predictions about the consciousness of artificial systems, or design non-conscious artificial systems that can be discarded without ethical concerns.

In the research meeting at the start of *A.I.*, a robot that genuinely loves its owner is presented as a major breakthrough. This is based on the old idea that AIs are rational rule-followers without feelings. In fact, robots with emotions are not as new as the film suggests. As our understanding of emotions has advanced, we have come to appreciate the important role that they play in human cognition, and this has led many researchers to build AI systems with emotions. Irreversible imprinting is a novel idea that has not yet been implemented. However, it is not clear why we would want this functionality, and the film highlights some of the problems that could occur with this type of system.

With the exception of Teddy, the mechas in *A.I.* are humanoid robots that are played by human actors. Real humanoid robots are complex mechanical systems that are expensive, unreliable, difficult to control and consume a lot of power. It is much easier to design robots for specific tasks - dishwashers are ubiquitous in our society; no-one has built a humanoid robot that can do the dishes. So it is very unlikely that the future will be populated with humanoid robots that directly replace human labor. Robot toys and companions are active research areas, particularly in societies with ageing populations, so we are likely to see increasingly sophisticated versions of Teddy in the future. However, the difficulties with humanoid robotics and human-level AI are likely to prevent us from building child robot companions like David for a long time.

A central theme of the film is David's desire to become real so that he can be loved by Monica. David could, in theory, become biological, but this would have many disadvantages. Without the appropriate history, this would be unlikely to help him achieve his central goal of being loved by Monica. *A.I.* does have good acting, particularly by Haley Joel Osment, who does an excellent job of portraying a child robot with a variety of odd behaviors (not blinking, rigid posture, inappropriate laughter, and so on). Mori's work on the uncanny valley nicely captures how David's not-quite human behavior in these early scenes produces an eerie lack of empathy. Teddy does not have this effect because he is clearly artificial, and Gigolo Joe is human enough to evoke empathy.

A.I. also raises issues about the displacement of humans by robots. In the film this occurs because of population control policies and a proliferation of robots. In the real world it is much cheaper and easier to breed self-repairing humans, than to manufacture and maintain humanoid robots to replace human labor. There are likely to be substantial job losses to AI in areas such as driving, call-centers, mining, and manufacturing. It is an open question whether the jobs created by AI will compensate for those that are lost. If AI does lead to mass unemployment, then it might be necessary to introduce a universal basic income.

References

Aldiss, B. W. (2001). *Supertoys Last All Summer Long: and Other Stories of Future Time.* London: Orbit.

Asimov, I. (1952). *I, Robot.* London: Grayson & Grayson.

Baker, C., Harlan, J. and Struthers, J. M. (2009). *A.I. Artificial Intelligence: From Stanley Kubrick to Steven Spielberg: The Vision Behind the Film.* London: Thames & Hudson.

Barrett, L. F. (2018). *How Emotions are Made: The Secret Life of the Brain.* London: Pan Books.

Berkeley, G. (1957). *A Treatise Concerning the Principles of Human Knowledge.* New York: Liberal Arts Press.

Breazeal, C. (2002). *Designing Sociable Robots.* Cambridge, Massachusetts and London: MIT Press.

Cangelosi, A. and Schlesinger, M. (2015). *Developmental Robotics: From Babies to Robots.* Cambridge Massachusetts: MIT Press.

Chollet, F. (2019) On the measure of intelligence. *arXiv*: 1911.01547.

Cleeremans, A. (2005). Computational correlates of consciousness. *Progress in Brain Research* 150: 81-98.

Collodi, C. (1995). *Pinocchio.* Ware: Wordsworth.

Crosby, M., Beyret, B. and Halina, M. (2019). The Animal-AI Olympics. *Nature Machine Intelligence* 1: 257.

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain.* New York: G.P. Putnam.

Gamez, D. (2007). *What We Can Never Know: Blindspots in Philosophy and Science.* London: Continuum.

Gamez, D. (2016). Are Information or Data Patterns Correlated with Consciousness? *Topoi* 35(1): 225-39.

Gamez, D. (2018). *Human and Machine Consciousness.* Cambridge: Open Book Publishers.

Gamez, D. (2021). Measuring Intelligence in Natural and Artificial Systems. *Journal of Artificial Intelligence and Consciousness* 8(2): 285-302.

Gardner, H. (2006). *Multiple Intelligences: New Horizons.* New York: Basic Books.

Grand, S. (2004). *Growing up with Lucy: How to Build an Android in Twenty Easy Steps.* London: Phoenix.

Gravato Marques, H. and Holland, O. (2009). Architectures for Functional Imagination. *Neurocomputing* 72(4-6): 743-59.

Haidt, J. (2013). *The Righteous Mind: Why Good People are Divided by Politics and Religion.* London: Penguin.

Hernández-Orallo, J. and Dowe, D. L. (2010). Measuring universal intelligence: Towards an anytime intelligence test. *Artificial Intelligence* 174: 1508-39.

Hess, E. H. (1958). Imprinting in Animals. *Scientific American* 198(3): 81-93.

Hingston, P. (2009). A Turing Test for Computer Game Bots. *IEEE Transactions on Computational Intelligence and AI In Games* 1(3): 169-86.

Hofstadter, D. R. (1979). *Gödel, Escher, Bach: an Eternal Golden Braid.* Hassocks: Harvester Press.

Husserl, E. (1960). *Cartesian Meditations: An Introduction to Phenomenology.* Translated by D. Cairns. The Hague: Martinus Nijhoff.

James, W. (2000). *The Principles of Psychology, Vol. 2.* New York: Dover.

Kätsyri, J., KlausFörger, Mäkäräinen, M. and TapioTakala (2015) A review of empirical evidence on different uncanny valley hypotheses: support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in Psychology | Cognitive Science*: 390.

Koch, C., Massimini, M., Boly, M. and Tononi, G. (2016). Neural correlates of consciousness: progress and problems. *Nat Rev Neurosci* 17(5): 307-21.

Lamy, L. (2015). Beyond Emotion: Love as an Encounter of Myth and Drive. *Emotion Review* 8(2): 97-107.

Legg, S. and Hutter, M. (2007a). A Collection of Definitions of Intelligence. *Proceedings of Advances in Artificial General Intelligence Concepts, Architectures and Algorithms: Proceedings of the AGI Workshop 2006*, edited by B. Goertzel and P. Wang, IOS Press, pp. 17-24.

Legg, S. and Hutter, M. (2007b). Universal intelligence: A definition of machine intelligence. *Minds and Machines* 17: 391-444.

Lorenz, K. Z. (1937). The Companion in the Bird's World. *The Auk* 54(3): 245-73.

Mori, M. (2012). The Uncanny Valley. *IEEE Robotics & Automation Magazine*: 98-100.

Oizumi, M., Albantakis, L. and Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. *PLoS Computational Biology* 10(5): e1003588.

Pessoa, L. (2019). Intelligent architectures for robotics: The merging of cognition and emotion. *Physics of Life Reviews* 31: 157-70.

Popper, K. R. (2002). *The Logic of Scientific Discovery.* London: Routledge.

Robertson, K. F., Smeets, S., Lubinski, D. and Benbow, C. P. (2010). Beyond the Threshold Hypothesis: Even Among the Gifted and Top Math/Science Graduate Students, Cognitive Abilities, Vocational Interests, and Lifestyle Preferences Matter for Career Choice, Performance, and Persistence. *Current Directions in Psychological Science* 19(6): 346-51.

Russell, S. J. (2019). *Human Compatible: AI and the Problem of Control.* USA: Penguin Random House.

Searle, J. R. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences* 3(3): 417-57.

Shanahan, M. (2008). A spiking neuron model of cortical broadcast and competition. *Consciousness and Cognition* 17(1): 288-303.

Shaw, R. C. and Schmelz, M. (2017). Cognitive test batteries in animal cognition research: evaluating the past, present and future of comparative psychometrics. *Animal Cognition* 20: 1003-18.

Teasdale, G. and Jennett, B. (1974). Assessment of coma and impaired consciousness. A practical scale. *Lancet* 2(7872): 81-4.

Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. *Biological Bulletin* 215(3): 216-42.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind* 59: 433-60.

Warwick, K. (2000). *QI: The Quest for Intelligence.* London: Piatkus.

Wilkes, K. V. (1988). ___, yìshì, duh, um, and consciousness. In *Consciousness in Contemporary Science*, edited by A. J. Marcel and E. Bisiach. Oxford: Clarendon Press, pp. 16-41.

Zeki, S. (2007). The neurobiology of love. *FEBS Letters* 581(14): 2575-9.