IN5480
Fall 2020

jrrasmus

Individual Assignment

Iteration 3

Table of Contents

1.1 Concepts, definition and history of AI and interaction with AI

1.1.1 History of AI

John McCarthy, an American mathematician and logician, was the person that coined the term *artificial intelligence* during a workshop in 1956. After World War II and up until this point the possibilities of computation had been theorized by great minds such as Alan Turing who discussed and published works regarding the possibilities of computers eventually obtaining human intelligence. During the late forties and early fifties conferences gathered researchers from many fields to discuss topics like neural network models and cybernetics. John McCarthy held one these conferences where he would put a name on the topic of discussion, *Artificial Intelligence*.[1]

1.1.2 Definition of AI

In an article posted at Stanford, November 12th 2007, John McCarthy shared his definition of *artificial intelligence* on a "layman's level" based on questions he often received from students, amongst others. To answer the question "What is artificial intelligence?", he wrote:

> "*It is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable*"[2]

The Encyclopedia Britannica defines artificial intelligence as such:

> *"Artificial intelligence (AI) is the ability of a computer or a robot controlled by a computer to do tasks that are usually done by humans because they require human intelligence and discernment. Although there are no AIs that can perform the wide variety of tasks an ordinary human can do, some AIs can match humans in specific tasks."*[3]

---

[1] Grudin, AI and HCI: Two Fields Divided by a Common Focus

[2] McCarthy, "What is AI", Stanford Education articles. 07.09.2020. http://jmc.stanford.edu/articles/whatisai.html

[3] Britannica, s.v. "artificial intelligence". 08.09.2020. https://www.britannica.com/technology/artificial-intelligence

The online dictionary Lexico, which is powered by Oxford, has this definition:

> *"The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages."*[4]

Based on aforementioned definitions, I conclude my own:

> *"It is the science around creating computational intelligence, where systems or machines have the ability to perform tasks similar or equal to human capabilities"*

This definition points out that artificial intelligence is about computer systems being developed to match human intelligence, and from there being able to approach and handle activities in ways a human would.

1.1.3 AI in Microsoft

Microsoft is deeply involved with artificial intelligence. Several of their products use AI, such as Office 365, Bing and Cortana, etc. On Microsoft's website they describe their approach "to responsible AI" where they believe AI can be used to help organizations achieve more. Following a list of AI principles, they take ethics of AI into account with the help of committees and bodies of offices like Office of Responsible AI (ORA). They talk about how AI can be used for good, working with environmental and humanitarian issues amongst other topics. Their own platform, Microsoft AI is presented as "robust framework" that can be applied to machine learning, data sciences or robotics development, among many other fields. Microsoft AI offers tools, infrastructure and services as well as training within the different fields AI can be applied.[5]

1.1.4 AI and Human Interaction in Fiction

The Netflix show Altered Carbon, which was based on a book series, has a very interesting portrayal of AI. The show takes place in the year 2384 where mankind has achieved "immortality" through the technology of *stacks*, discs that contains a copy of your consciousness. In the futuristic metropolis Bay City, the streets are filled with so-called "AI-hotels". These hotels are completely run with AI-hosts that manifests in physical holograms (made of nanobots?) which can appear anywhere around and in their

---

[4] Lexico, sv. "artificial intelligence". 08.09.2020. https://www.lexico.com/definition/artificial_intelligence
[5] Microsoft, "AI Platform".

hotel. The AI essentially is the hotel and can change the building's physical appearance based on their customers' needs and wants. In the year the show is set, it seems that AI hotels are a bit outdated as the protagonist is told that "nobody stays in them anymore" and that they are worse than an "over attached girlfriend" (paraphrased). In this world, all your personal information is connected to and can be accessed through your DNA, so in order to use the services of the hotel you can for example give your fingerprint. Once the transaction is done the AI will protect and assist its guests at all costs. The AI in this case does indeed become very attached and appear very human in the process (looks, actions, conversations etc.). Another interesting detail in the show is how the AI hotels have meetings together where they discuss how some of them take advantage of humans or how others are too fond of them.

1.2 Robots and AI systems

1.2.1 Origin of the Word Robot

Karel Čapek, a Czechoslovakian writer and journalist introduced the phrase *robot* in the drama *Rossum's Universal Robots*. The story was a protest against the uprise of modern technology, in which he writes about artificial humans who evolves and turns on mankind. The word derives from the Czech word *robota*, which means serf or laborer. [6]

1.2.2 Definition of Robot

The International Organization for Standardization uses this definition for industrial robots:

> *"automatically controlled, reprogrammable, multipurpose manipulator, programmable in three or more axes, which can be either fixed in place or mobile for use in industrial automation applications."*[7]

This is quite a mechanical definition of a robot as it tells us how, albeit briefly, it should be able to move and how it can be manipulated in order to be defined as a robot.

Oxford Learner's Dictionaries defines the word robot as such:

> *1. A machine that can perform a complicated series of tasks automatically*
>
> *2. (especially in stories) a machine that is made to look like a human and that can do some things that a human can do.*[8]

---

[6] Store norske leksikon, s.v. " Karel Čapek ". https://snl.no/Karel_%C4%8Capek
[7] International Federation of Robotics. "Standardization".
[8] Oxford learners dictionaries, s.v. "Robot".
https://www.oxfordlearnersdictionaries.com/definition/american_english/robot

These definitions highlights that the robots ability to perform actions or tasks on its own accord, just as a human would, is what makes it a robot.

Based on these definitions I have concluded my own:

> "*An autonomous machine, which through preprogrammed physical movements, can perform tasks in accordance with human-like abilities*"

In order to be defined as a robot, a machine must be able to move and perform tasks with a certain grade of independence. In order to do so it must be programmed and built in a way which allows this, hence the comparison to humans.

1.2.3 Relation Between AI and Robots

AI and robots are closely connected. Nearly all aforementioned definitions highlight the human-like attribute for both AI and robots. What separates them, from my point of view, is that a robot acts within the limits of its programming, which often entails a range of preset actions, whereas AI can expand and develop these actions through learning and research. Another difference is that AI can be applied to a wide variety of computational systems or machines, whereas robots require a physical and mechanical manifestation. See for example the definition by John McCarthy vs. the definition by ISO. Furthermore, it would seem that all robots are implemented with AI, while not all AI constitutes as robots. However, it makes you wonder, must a robot use AI in order to be classified a robot? If I program a robot-arm to grab an object like a human would, with some simple lines of code and without the use of AI, would it still be a robot? Based on Oxford's definition, perhaps not? Depending on the grade of independence at which this arm performs it's task. E.g. would I have to press "start" when placing an object in front of the arm, or would the arm sense the objects presence on its own and then grasp it?

1.2.4 Example of a Contemporary Physical Robot

Boston Dynamics is a company that design and produce mobile robots. One of their products is SPOT, a medium-sized, four-legged yellow robot that can move through tough terrain, climb stairs and even get back on its "feet" after falling. SPOT can be used in a variety of contexts and for different purposes, such as construction, mining, healthcare and entertainment, just to mention a few. SPOT is equipped with several sensors and customizable software that helps it read its surroundings and perform tasks. SPOT's actions and movement can be preprogrammed but he can also be controlled with a controller. The

controller has buttons for moving SPOT in different directions and to grab things etc. It also has a display which shows the views from the several cameras installed in SPOT.[9]

Take a look at SPOT's launch video or Adam Savage's take on the little robot!

https://www.youtube.com/watch?v=wlkCQXHEgjA, https://www.youtube.com/watch?v=R-PdPtqw78k

1.3 Universal Design and AI systems

1.3.1 Definition of Universal Design

The definition of universal design, originating from The Disability Act 2005, and posted by The Centre for Excellence in Universal Design, reads as follows:

1. *"The design and composition of an environment so that it may be accessed, understood and used*
    i. *To the greatest possible extent*
    ii. *In the most independent and natural manner possible*
    iii. *In the widest possible range of situations*
    iv. *Without the need for adaptation, modification, assistive devices or specialised solutions, by any persons of any age or size or having any particular physical, sensory, mental health or intellectual ability or disability, and*
2. *Means, in relation to electronic systems, any electronics-based process of creating products, services or systems so that they may be used by any person"*[10]

It means that electronic information systems should be available for their proper and intended use by any person, regardless of any ability or disability they might have. The purpose is to secure equal inclusion of all people who might use a certain service or product. It also avoids exclusion where some people may be unable to use the service or might require special adaptations in order to use it.

1.3.2 Potential of AI with Respect to Human Abilities

The possibilities are vast for the potential of AI when it comes to patching or extending human capabilities. One example is Elon Musk's company Neuralink that is working on an AI chip that connects to your brain. If successful, the chip could be able to give people with robot limbs the sense of touch or

[9] Boston Dynamics. "SPOT"
[10] Universal Design. "Definition and overview"

be able to treat illnesses like Parkinson's.[11] Another example comes from École polytechnique fédérale de Lausanne where they are developing an artificially intelligent robotic hand. With machine learning the hand allows the bearer to control each individual finger by learning their movements.[12]

Considering these examples and the rapid speed in which technology is evolving, we might be able to fill all the gaps in human functionality. Whether you are impaired physically or cognitively won't hold you back in society as enhancements will be able to replace, repair or even improve what you lack.

1.3.3 Potential of AI with Respect to Inclusion and Exclusion

All AI models are affected by bias, whether it stems from the data scientists, data engineers or the data itself. Depending on the purpose of the model it can have negative or positive consequences. A well-known example is the AI system COMPAS which wrongly predicted a resurrection in crime of known African American perpetrators.[13] On a positive note, AI can also bring people together. DeepL is a company that builds AI translation systems. They aim to "break down the language barriers worldwide and bring cultures closer together".[14]

1.3.4 Concept of Understanding

I would say, that when understanding something you can make correct assumptions and interpretations of something, perhaps new or unfamiliar, based on cognitive abilities and previous experiences. I think machines can "understand" what we program them to understand. We give it sets of data and a recipe for how to interpret it, this is how it "understands". But without this foundation it wouldn't be able to make sense of something entirely new to it the same way a human would.

---

[11] Hamilton, "Elon Musk's AI brain chip company Neuralink is doing its first live tech demo on Friday. Here's what we know so far about the wild science behind it". Business Insider. https://www.businessinsider.com/we-spoke-to-2-neuroscientists-about-how-exciting-elon-musks-neuralink-really-is-2019-9?r=US&IR=T

[12] Carfagno, "AI-Powered Prostethic Hand Provides Unprecedented Control for Amputees". Docwirenews. https://www.docwirenews.com/future-of-medicine/ai-powered-prosthetic-hand-provides-unprecedented-control-for-amputees/

[13] McKenna, "Three notable example of AI bias". AI Business. https://aibusiness.com/document.asp?doc_id=761095&site=aibusiness

[14] DeepL. "Another breakthrough in AI translation Quality". https://www.deepl.com/blog/20200206.html

1.4 Guideline for Human-AI interaction

The fourth guideline, *Show contextually relevant information,* entails that the AI steps in with relevant information to the user at appropriate times. An example of this a conversation I've had with Google Home. When asking Google Home "How nutritious are mangoes?" it tells me which page it collects the information from, reads me a snippet and then asks if I want "more context". Another example is when I ask Google Home to convert a unit for me, for example from pounds to grams. It then first tells me the conversion and after, how to calculate it myself ("divide with approximately 2.2").

When looking at the User Interface Design Guidelines: 10 Rules of Thumb from the Interaction Design Foundation, several similarities can be found. Both sets of guidelines highlights the importance of showing the user what the system is doing and making its abilities clear. Both points out how the system should match concepts between the system and the real world, making it more predictable and easier to use. Assisting users with errors and backtracking is another point they have in common. A difference between the two could be that some of the design guidelines focus a bit more on the visual design, e.g. keeping the systems "aesthetic and minimalistic". Furthermore, some of the Human-AI Interaction guidelines highlights other aspects such as the importance of avoiding social biases and how the system can learn from the user and adapt over time.[15]

2.1 Characteristics of AI-infused systems

2.1.1 Key characteristics of AI-infused systems

In the lecture regarding AI-interaction from the second module, the characteristics of AI-infused systems are listed as follows:

The characteristic *learning* entails that the AI dynamically adapts based on what it learns through interaction with users over time. This is useful as it allows a user's experience to be personalized when interacting with AI.

[15] Wong, "User Interface Design Guidelines: 10 Rules of Thumb". https://www.interaction-design.org/literature/article/user-interface-design-guidelines-10-rules-of-thumb

As the AI develops through interaction, mistakes are not uncommon , therefore *improving* is a key characteristic. By *improving* the AI through having users verifying it's suggested actions, we can avoid possible dangerous, unjustified or costly actions and consequences.

Black box entails that whatever happens "behind the scenes", how input is handled and how the output is made, is not visible to the users. When this is information is not available to the users, it is important to clarify the systems actions.

For the AI to learn and adapt it needs data, therefore another characteristic is *Fueled by large data sets*. Data is given to the AI or collected by the AI through the interaction with users.[16] [17]

2.1.2 Example of AI-infused system

As an example of AI-infused system I choose Spotify. Through my experience, Spotify portrays several of the characteristics quite well, it collects a lot of data regarding the music I listen too, what I listen too often and what music I have liked and saved. Based on this it has learned my taste in music and gives good recommendations through the *discover weekly* playlist (albeit, it is not always great). Based on several of my playlists it suggests new mixed playlists with similar genres or general composition. I have no idea how the Spotify AI does what it does, but it's actions and the results of those actions are clear. Of course, it is not always perfect. Sometimes *Discover Weekly* lists songs that I already have saved from before.

How Spotify performs these actions through the different characteristics has a positive effect on the user experience, in my opinion. It can be hard to find new music sometimes and to find inspiration for new playlists. I look forward to turning on Spotify as I know that I will gain new (music) insight when using it. I don't feel the need to understand what is going on inside the black box as the output usually is satisfactory.

---

[16], Følstad , Asbjørn, "Interaction with AI – Module 2", (Lecture Notes, UiO, September 22, 2020)
[17] Amershi, et al. "Guidelines for human-AI interaction"

2.2. Human-AI interaction design

2.2.1 Main take-aways from Amershi et al. (2019) and Kocielnik et al. (2019)

Amershi et al. point out how AI poses new challenges and opportunities when it comes to designing user interfaces. Although different features such as speech recognition, face recognition, translation, object recognition, and so on, are developed and further improved, they are still not perfect and often result in false positives and negatives. Furthermore, behavior that can be viewed as offensive, confusing or dangerous may emerge from the unpredictability of AI-infused systems. There is a risk that existing usability guidelines, when it comes to common user interface design, won't be upheld with the use of AI. To remedy this, the authors analyzed 20 years of work regarding AI design in order to create a set of guidelines for human-AI interaction. This resulted in 18 guidelines, grouped in four different categories depending on when the interaction takes place (*initially, during interation, when wrong, over time*). The guidelines were tested, refined and reviewed through four iterations by researchers, designers and usability practitioners.[18]

Kocielnik et al. points out the importance of user expectations and how this can affect the user experience when interacting with AI systems. Different users will have different, and often very high, expectations of the usability and capabilities of an AI-system. If these expectations aren't met, it can have a negative impact as users may end up disappointed, unsatisfied with the product and less willing to use it again. With AI-infused systems, new functionalities have emerged that further affects the user experience of a product, followed by new performance expectations regarding these features. *Natural language understanding, sensor-based inferences, object recognition in video or images* are a few examples of such functionalities. These are described as probabilistic and "*almost always operating at less than perfect accuracy*" in the article, which likely doesn't fall in line with the users' expectations of a consistent and perfect product. They mention that these expectations can be shaped by the information that is shared about the system, what knowledge and understanding the user already possesses, and their previous firsthand experiences.

Knowing the effect expectations can have on the user experience with AI-powered technologies, Kocielnik et al. explored three different techniques that could help form these expectations, as well as help them further their understanding of how this affects user acceptance. The first technique

[18] Amershi, et al. "Guidelines for human-AI interaction"

constituted of an *Accuracy Indicator*, which informs of the system accuracy. The second is *Examples based Explanation*, which aims to expand user understanding. Third is *Performance Control* where the user is given power to adjust the system performance directly. These techniques where used with a Scheduling Assistant, and AI-powered system. Two versions of the systems were tested, one in which False positives where avoided, and in the other, False Negatives.

They concluded that their techniques indeed, efficiently had a positive impact on user expectations and experiences of the Scheduling Assistant. Furthermore, they found that putting less focus on a system that commits more False Positives mistakes (High Recall) instead of False Negatives (High Precision) can result in lowered acceptance and significantly decreased accuracy perception. Their findings show that this can be used in order to better "user acceptance of AI technologies".[19]

2.2.2. Discussion of guidelines in Amershi et al. (2019)

Spotify somewhat adheres to guideline 5, *Match relevant social norms*. It is not an AI system you converse with in anyway and is not supposed to have a human representation. However, music can be a very cultural thing and different genres of music or artists will be more prevalent in different parts of the world. Spotify suggests a lot of Norwegian artists on the frontpage, which is relevant since I live in Norway. Further, it suggests popular artists, albums or playlist based on what is currently very popular.

Guideline 13 *Learn from user behavior* is also followed, as mentioned in assignment 2.1.2., Spotify notices what you listen to over the period of one week, then creates a playlist that is inspired by all the music, regardless of genre or artist. I definitely have noticed that this function has become better. A few years ago, I would mostly be disappointed in the music that was suggested, however now, I enjoy the new suggestions more often than not. Another aspect I like is that if I turn on *private session*, Spotify does not take these songs into account when creating the *discover weekly* playlist.

I am often not interested in the suggestions on the frontpage (local artists, current popular music) so it could be an improvement to, through a combination of these two guidelines, make it so that Spotify better understands what genres I like and suggest popular artists/music based on that. Or based on what some of my friends are listening to.

[19] Kocielnik, "Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI systems"

2.3 Chatbots / conversational user interfaces

2.3.1 Key challenges in the design of chatbots / conversational user interfaces

One of the key challenges when designing chatbots is how the approach is different compared to traditional interaction design, where visual layouts in the shape of graphical user interfaces and visual interaction mechanisms is used when designing for usability. Now, however, these useful skills will be less in demand as the design object is conversation itself. Følstad and Brandtzæg writes "*We need to move from seeing design as an explanatory task to an interpretational task*", which means that the focus need to shift from clarifying the user's possibilities, to instead understand the users and their needs.

The focus within HCI research has been on the interactive system itself, i.e. the design object, and not as much on the goal of the users, and with conversational user interfaces being on the uprise, the focus may need to change. The necessity to move from UI design to service design is another key challenge, as different sources, services and contents, that were previously separated, will "*blur into the same conversational threads*".

The necessity to design for human and AI interaction in networks is another key challenge. In traditional interaction design the focus is often on one device and/or one user, but with conversational user interfaces, such as chatbots, multiple actors can be part of the interaction. This may cause many unpredictable and possibly unredeemable consequences, as seen in the example of Microsoft's chatbot Tay.[20]

A proper understanding of user needs and how to capture the user experience with the help of service design as well as how to safely connect multiple actors in the interaction, may lead way to a very interesting future with marvelous possibilities for human-AI interaction.

2.3.2 Key challenges in the design of chatbots and guidelines G1 and G2 in Amershi et al. (2019)

Guideline 1, *make clear what the system can do*, can be helpful when designing for usability in conversational user interfaces when designers no longer can rely on the visual layout and graphical interfaces. Even though user's can't see menus, icons and the like, it needs to be clear what options the users have when using conversational user interfaces. The second guideline, *make clear how well the*

[20] Følstad & Brandtzæg, "Chatbots and the new world of HCI"

system can do what it can do, is also relevant as graphical feedback in the shape of error messages or information pages may disappear in future conversational user interfaces. By following these guidelines, you can further meet the needs and expectations of the users.

3.1 Human AI collaboration

3.1.1 Big Dog

Big dog is a dog or mule like military robot designed to assist soldiers in carrying cargo through rough terrain.[21] The robot is unmanned and have a variety of sensors related to for example joint position, force, gyroscope and temperature, which help guides its path. Big Dogs actions, sensor management and communication, among other things, are controlled through it's computer, which a human operator access remotely. Through the operator control unit, IP radios and a visual display, the human driver can adjust direction and speed of the robot, tell it to stop and go, amongst several other functions. However, operation of the legs, adjusting itself on difficult terrain and to unexpected disruptions from the environment, are run on the onboard control system.[22]

Looking at Sheridan and Verplank's ten levels of automation/autonomy, Big Dog could be on level 7; *the computer executes automatically, then necessarily informs the human*. Big Dogs movement and adjustment to terrain is automated, however *what* and *how* it performs (walk/run, stop/go, sit/stand, etc.) can be controlled by a remote human operator. I think this description best matches that of level 7 but am not sure if this level best describes Big Dog's level of autonomy. For example, it is not entirely clear what "necessarily informs the human" entails regarding Big Dog, and whether the operator constantly observes and take over control when necessary or if they wait for cues from the robot. The fact that Big Dog is independent and automated in its main tasks but also can be controlled by human operator when necessary, puts it in the top right quadrant of the two-dimensional framework of Trusted, Reliable & Safe. The top right quadrant of the framework entails tasks that are complex, hard to grasp and occurs in a variety of context[23], which I think matches the described purpose and use of Big Dog in a military setting.

---

[21] Philips, E, et al. "Human-animal teams as an analog for future human-robot teams"
[22] Raibert, et al. "BigDog, the Rough-Terrain Quaduped Robot"
[23] Schneiderman, B., "Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy AI"

If Big Dog's level of autonomy was increased, it could lead to benefits where operators or other staff could spend less time to monitor the robot and more time on other useful tasks. This could also be beneficial in a critical military setting where Big Dog could just act quickly, judging it's own surroundings, instead of an operator taking control remotely, trying to navigate and make decisions based on the data received through the different sensors. However, this would require that involved humans (e.g. soldiers) fully could trust and rely on the robots independent abilities and feel safe regarding its actions, something that could be difficult to achieve. It is perhaps more likely that in a situation where lives are at stake, a human with the proper military and combat experience would prefer to take control despite the robot's level of autonomy. A decreased level of autonomy would require more human interference, and therefore also more staff that are in charge of the robot, as well as more resources (training of control staff and maintenance staff). If a lower level of autonomy would entail direct control of the robot, being in its vicinity (i.e. no remote control), it could lead to a disadvantage if for example one soldier needs to control and monitor the robots at all times, requiring their attention and making them less aware of their, potentially dangerous surroundings. However, with manned control, perhaps the soldiers would feel safer knowing that one of their own with the proper experience took the reins out in the field.

Without detailed knowledge regarding how Big Dog is used or communicated with and how the interface or controls look, it is a bit difficult to properly asses it's explainability. While the robot moves around independently, it sends a lot of data through its various sensors to the remote operator, whom interpret and take action/control of Big Dog based on this data. In order for this to be possible, the operator needs to be able to understand and interpret this data, therefore I argue that this AI to at least some extent is transparent. If the cause of Big Dog's actions and the data from the sensors are explained to the operator, causality has also been taken into account. As this robot adapts to and moves in real and tangible surroundings, I would say there is little bias present in how it interprets the world. However, if Big Dog were supposed to detect potential threats, that would be a different matter as different people would categorize different threats (one nations military vs. another). For the same reasons, I'd say this AI is fair. Regarding it's autonomous tasks, I would also say Big Dog appears to be safe and reliable. You don't need a detailed explanation of

why the robot all of sudden changed course, you can see that the boulders were a bit too big to walk across, so the robot had to go around instead.[24]

3.1.2 Paro

Paro[25] is a therapeutic robot, resembling a cute seal and designed with the purpose of keeping elderly people company and reduce depression (Schneiderman). The robot may help reduce stress and improve relaxation and motivation with it's users. It can improve socialization between users and stimulate the patient-caregiver interaction. Paro does all of this with the help of five different sensors (tactile, light, audition, temperature, posture), which allows it to beware of its surroundings and people, to differentiate between light and dark and to recognize different voices and phrases. Paro can sense being touched, and this ability is used to train his behavior. The user can stroke the seal to reinforce positive behavior or hit him to teach him what actions not to do. The robot appears alive, responding to different stimulus, moving it's body and make different sounds.[26]

Based on Sheridan and Verplank's ten levels of automation/autonomy, Paro could be on a level 8; *the computer informs the human only if* asked. The robot interprets and acts autonomously based on it's surroundings with the help of it's different sensors. How it does this is preprogrammed and not determined by human control in the context of use. However, it is not entirely autonomous as the user can affect it's behavior through different actions, which I think lines up nicely to the description of level 8 (maybe also level 7). Looking at the Trusted, Reliable & Safe framework, I would also place Paro on the top right quadrant as the robot has a high level of automation but some of it's behavior can be modified by a human. Furthermore, it doesn't perform any tasks that could be damaging or life threatening as it's purpose is to have positive psychological effects on the user through interaction. It's animal like appearance induces trust, and can therefore be considered to be trusted, reliable and safe.

As a companion robot which behaves much like an animal, an increased level of autonomy in Paro could result in disadvantages. For example, a real pet would have a unique personality and would be trained and shaped differently by different owners, whose' preferences and habits could vary greatly. If Paro was so autonomous that the user couldn't change its behavior, it could lead to Paro being compatible

[24] Hagras, H. "Toward Human-Understandable, Explainable AI"
[25] [25] Philips, E, et al. "Human-animal teams as an analog for future human-robot teams"
[26] http://www.parorobots.com

with fewer users as not all would respond positively to it's standardized behavior. If Paro is going to have a positive psychological effect on the user, I believe it should be moldable in way similar to a real pet, where the owner can reinforce or discourage different behaviors. It could potentially also lead to confusion as the user might not always understand why it acts the way it does. Making Paro less automated could still provide comfort (just as a regular plush animal could) and perhaps make the users feel more in control, for example if they could give commands ("act as if you want a cuddle") or adjust its behavior (only communicate through movements and not sound). This, however, might go against Paros purpose as a pet-like companion robot.

Considering the purpose and use of Paro, this AI should be considered explainable. The robot acts much like an animal, and communicates in similar ways to a real-life pet, which can be understood by its target group, therefor I think transparency and causality is present in this AI. Paros programmed behavior and actions could possible be affected by bias from it's creators, stemming from their view on what kind of behavior would have a positive effect on mental health. Whether Paro is fair and safe, relies on us being able to trust that it behaves as expected and allows humans to make adjustments when necessary, which I believe it does, based on the description of Paro.

References

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Teevan, J. (2019). Guidelines for human-AI interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (paper no. 3). ACM.

Boston Dynamics. "SPOT". 06.09.2020. https://www.bostondynamics.com/spot

Britannica, s.v. "artificial intelligence". 08.09.20. https://www.britannica.com/technology/artificial-intelligence

Carfagno, "AI-Powered Prostethic Hand Provides Unprecedented Control for Amputees". Docwirenews. 12.09.2019. https://www.docwirenews.com/future-of-medicine/ai-powered-prosthetic-hand-provides-unprecedented-control-for-amputees/

DeepL. "Another breakthrough in AI translation Quality". 06.02.2020. https://www.deepl.com/blog/20200206.html

Følstad , Asbjørn, "Interaction with AI – Module 2", (Lecture Notes, UiO, September 22, 2020).

Følstad, A., & Brandtzæg, P. B. (2017). Chatbots and the new world of HCI. interactions, 24(4), 38-42.

Grudin, Jonathan. 2009. “AI and HCI: Two Fields Divided by a Common Focus”. *AI Magazine* 30 (4), 48. https://doi.org/10.1609/aimag.v30i4.2271.

Hagras, H., Toward Human-Understandable, Explainable AI, Computer, 51, 9, 2018, 28-36 https://ieeexplore.ieee.org/document/8481251

Hamilton, "Elon Musk's AI brain chip company Neuralink is doing its first live tech demo on Friday. Here's what we know so far about the wild science behind it". Business Insider. 26.08.2020. https://www.businessinsider.com/we-spoke-to-2-neuroscientists-about-how-exciting-elon-musks-neuralink-really-is-2019-9?r=US&IR=T

International Federation of Robotics. "Standardization". 09.09.2020. https://ifr.org/standardisation

Oxford learners dictionaries, s.v. "Robot". https://www.oxfordlearnersdictionaries.com/definition/american_english/robot

Kocielnik, R., Amershi, S., & Bennett, P. N. (2019). Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (paper no. 411). ACM

Luger, E., & Sellen, A. (2016). Like having a really bad PA: the gulf between user expectation and experience of conversational agents. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 5286-5297). ACM

Lexico, sv. "artificial intelligence". 08.09.2020. https://www.lexico.com/definition/artificial_intelligence

McCarthy, "What is AI", Stanford Education articles. 07.09.20. http://jmc.stanford.edu/articles/whatisai.html

Microsoft, "AI Platform". 08.09.2020. https://www.microsoft.com/en-us/ai/ai-platform

McKenna, "Three notable example of AI bias". AI Business. 14.10.2019. https://aibusiness.com/document.asp?doc_id=761095&site=aibusiness

Paro. "Paro Therapeutic Robot". Information gathered 12.11.2020. http://www.parorobots.com

Phillips, E., Ososky, S., Swigert, B. and Jentsch, F. Human-animal teams as an analog for future human-robot teams, Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol 56, Issue 1, (2016) pp. 1553 – 1557 DOI: https://doi.org/10.1177/1071181312561309

Raibert, Marc & Blankespoor, Kevin & Nelson, Gabriel & Playter, Rob. (2011). BigDog, the Rough-Terrain Quadruped Robot. Proceedings of the 17th World Congress. Vol 17.

Store norske leksikon, s.v. " Karel Čapek ". https://snl.no/Karel_%C4%8Capek

Shneiderman, B., Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy, arXiv.org (February 23, 2020). https://arxiv.org/abs/2002.04087v1 (Extract from forthcoming book by the same title)

Universal Design. "Definition and overview. 08.09.2020 "http://universaldesign.ie/What-is-Universal-Design/Definition-and-Overview/

Wong, "User Interface Design Guidelines: 10 Rules of Thumb". Interaction Design Foundation. 08.09.2020 https://www.interaction-design.org/literature/article/user-interface-design-guidelines-10-rules-of-thumb

Appendix 1: Feedback from first iteration

After the first iteration I received useful feedback which I took into account when starting with the second iteration. My arguments and descriptions of the different definitions, as well as use of references was received as positive. Therefore, I aimed to maintain my referencing approach and make sure I have well written descriptions and arguments.

There was however room for improvement in the structure of the text, as it wasn't clearly listed which questions were answered in the different paragraphs. I was advised to add the numbers of the different questions, or the questions themselves, in the text. I took this to heart and added a table of contexts and included all the different questions, including their numbers, to make it easier to navigate through the text. To make it even more neat and structured, I also added a front page and made sure that the new assignments from the second iteration follows the first iteration.

Furthermore, regarding the discussion around the relation between AI and Robots, there was an interesting and important point that I had failed to mention; the fact that AI can "learn" based on it's programming, and how this is one the differences between AI and robots. Robots are usually programmed to perform a set of actions, but with the help of AI, they can "learn" to perform different actions based on the research they can do themselves. I definitely agree on this comment from my fellow student, so made sure to add something regarding this fact, to my original answer.

Appendix 2: Feedback from second iteration

After the second iteration I received two stars and one wish. My fellow student commented that I wrote good and detailed summaries regarding the article by Kocielnik et al. It was also mentioned the I exemplified Spotify in a good way, and that it was clear to see the connection between the different parts of the assignment, as well as how Spotify constitutes as an AI-induced system. This was a very nice feedback and I will try to main the same level of quality throughout the third and last iteration.

The wish consisted of a desire to read more about the different topics mentioned in section 2.3. My fellow student commented that they found these themes interesting but speculated that adding to the section might require some work. This was also good feedback, however with limited time and no clear instruction on what exactly to write, I only added a few sentences.