Helge Moes
11348801
Human(e) AI
5: Do you trust AI?
Dieuwke Hupkes & Jaap Jumelet
Assignment 6: Final Essay
Word count: 1910

Who to trust; artificial intelligence or humanity?

Introduction

Artificial Intelligence has been integrated in not only society, but also increasingly closer to humans. It has become an integral part of how people function, namely decision-making of human beings is delegated through automated processes (Araujo et al., 2020). For instance, from generating information to automatically parking cars, artificial intelligence leaves less to be desired. Nevertheless, humans act and can be regarded as the sources for the data that forms artificial intelligence. Collected data can even be used for political benefit and target certain groups of the society and penetrate their privacy and in turn influence their political opinion (Dobber et al., 2017). Nowadays, this is done by social media and the archival role it plays, which in this case allows for information to generate ads that are created to the preferences of the individual user (Leerssen et al., 2019). Furthermore, people use artificial intelligence as to benefit themselves with the information, which in turn leads to the questioning of ethics and morals of how artificial intelligence is used to manipulate humanity. It can be seen as a mirror that reflects humans' behavior. Therefore, with regard to the question whether artificial intelligence is to be trusted, I would rather question whether humans are to be trusted.

To investigate the notion whether artificial intelligence or humanity is to be trusted, the research will not only consider journal articles and academic papers, but also movies that have portrayed domains where artificial intelligence takes over the human's ability of choice. Movies such as: *Her* (Jonze 2014) *I, Robot* (Proyas 2004) and *2001: A Space Odyssey* (Kubrick 1968), portray a society where artificial intelligence has been integrated to such a degree that it is a fundamental part of households and how humans operate. These

movies made assumptions of how artificial intelligence plays an integral role and how it demobilizes or enhances the human's capability to function, for example humans do not need to drive cars, do grocery shopping, clean the house, take care of the children etc. Eventually, artificial intelligence takes over control and a shift occur in these movies, which leads to the demise of humankind. The asymmetry of roles becomes difficult to determine and it is challenging to trace back who is accountable for this outcome, humans or artificial intelligence. However, this article will focus mainly on the domain of households that will be run by artificial intelligence and the requirements and repercussions this will have on society as we know it through examining different examples.

Methodology

The research methodology can be characterized as a content analysis that also incorporates movies as a subject of reflection. Furthermore, the arguments that will be assessed are that artificial intelligence is to be trusted, since technology and humans have been collaborating on producing societal development. However, the research will also criticize artificial intelligence, since it is still developing and therefore difficult to trust, since it contains flaws in its current state. Finally, by assessing the movies that show the function artificial intelligence, the argument of a doomsday scenario that involves artificial intelligence, does not have to be a self-fulfilling prophecy.

Trust in artificial intelligence

Artificial intelligence can be described as a computer system that is able to carry out tasks that are normally performed by human intelligence. In first glance, artificial intelligence is to be trusted, since it has been a part of humanity for many years and therefore an integral part of us as humans. There is a reason for algorithmic appreciation is being researched. Technology has become so advanced, that it contains personal characteristics that allow automated decision-making to be done based on a more 'human' fashion (Araujo et al., 2020). Sherry Turkle noticed how people would perceive computational creatures as "relational artifacts" (Turkle 2007 501), such as humans that are emotional machines.

However, artificial intelligence still is an entity that functions on rationality such as algorithms. Since we cannot physically see how artificial intelligence operates, we treat it as a mind of its own. This example is portrayed in Spike Jonze's *Her* (Jonze 2014). The protagonist, Theodore, falls in love with his computer

and builds a relationship with it. The operating system, OS1, meets every desire of Theodore, but it is not certain whether this is done out of emotion or because the artificial intelligence is programmed to do so. In this case, artificial intelligence shows a huge potential to become a 'relational artifact' and aide humanity with its emotional challenges. It becomes part of humanity and can provide humanity a sense of interaction, which many people are not mentally able to, such as severely autistic or depressed individuals like Theodore.

Who is to be trusted?

Despite the overwhelming opinion not to trust artificial intelligence, since it is difficult to understand how it operates, the results of automated decision-making exceeded that of humans within specific situations (Araujo et al., 2020). Therefore, one might argue that trusting artificial intelligence is an individual matter. This also is conceptualized within Alex Proyas' *I, Robot*, where robots are able to function as humans' slaves. Initially this seems to be valid, because the robots are perceived as not containing emotions and enabled to process their own thoughts, yet in practical sense, these robots can perform practical matters better than the humans that are dependent of them and are able to develop emotions. Artificial intelligence is used as a solution to the humans' incompetence that has gradually developed through evolution and culture.

Additionally, a framework to guide ethical governance in robotics and artificial intelligence is fundamental to gain the trust of humanity (Winfield & Jirotka 2018). That what obstructs the trust, is that humanity is exposed to disruptive new technologies that are challenging to comprehend. In this case, trust can be created by transparency, inclusiveness and agile ethical governance (Winfield & Jirotka 2018) from the manufacturers that built these contraptions. An explanation of artificial intelligence has to contain human features, such as philosophy, psychology and cognitive science for people to understand such an abstract entity (Miller 2019). For this reason, Alex Proyas portrayed robots as humans that live and experience emotions, like us humans do. It holds a potential that keeps enriching humanity and eventually becomes part of society. This not only proves how close technology is to humans, but it is also a reason for it to be trusted.

Nevertheless, movies such as *I, Robot* and *2001: A Space Odyssey* also show that artificial intelligence holds a potential to overtake humanity. Therefore, artificial intelligence is to be criticized, as it is a developing entity that means it can perform beneficial for humanity or become a threat. In Stanley Kubrick's masterpiece

that is *2001: A Space Odyssey,* HAL 9000 is the on-board computer of the spaceship the Discovery One (Kubrick 1968). The advanced technology is produced to assist humanity to discover the mysterious monolith. However, due to a malfunction, HAL 9000 starts eradicating the crew one by one. This notion of humanity being victimized due to their imperfection, evokes a judgmental overtake by artificial intelligence, which also happens similarly in *I, Robot.*

Technology, for instance HAL 9000, has been currently developed, since even chatbots can interact with humans and even replace humans in fields as customer-service and online tutoring (Go and Sundar 2019). Anthropomorphic cues allow for artificial intelligence to interact with humans, this is also displayed in *2001: A Space Odyssey,* when HAL 9000 is able to lipread the humans on the ship that are discussing how to take out the onboard-computer. Furthermore, this technology is still to be realized, but will be developed in the foreseeable future.

An example of such technology is the In-home Voice Assistant; Alexa. This piece of human creation is focused on voice interaction with human beings. Where HAL 9000 is still ahead of current technological capabilities, findings of Mclean and Osei-Frimpong prove that Alexa generated utilitarian, symbolic and social benefits for individuals (McLean and Osei-Frimpong 2019). However, such advanced technology also provides for problems, for instance privacy risks affect the use of in-home voice assistants. Alexa might be hacked and it will allow for third parties to follow what happens within a household. In this case, the fear of a HAL 9000 is still visible when it comes to digital development.

Artificial intelligence as a source of doubt

For some reason, artificial intelligence perceives humanity as a hazard that has to be eliminated. Yet, if we consider that humanity and technology are both two agents, there is a risk that one of these to be malevolent and ill-willed (Marsh 2005). In this case, if we perceive technology to reflect humanity, a black mirror so to speak, then humanity is not to be trusted, since the use of technology contained underlying intentions. As by taking the risk to trust technology, thus minimizes the potential damage humanity may cause (Marsh 2005). There is a symbiosis of trust in technology, both these agents react to one another, but it is the communication between both parties that causes for distrust (Hengstler et al., 2016). This also shows that both entities need one another to function properly and that trust is achieved when transparency realized.

When we translate what the movies portray about artificial technology to a scenario that plays nowadays, we observe how political campaigns use data to target voters in *Two Crates of Beer and 40 Pizzas: The Adoption of Innovative Political Behavioural Targeting Techniques* (Dobber et al., 2017). In this case, it is the ill intentions of the human that grants artificial intelligence the rights to breech privacy and manipulate others by ads. These antics are not conceived as unethical, since there are no specific data regulations that restrict such practice (Dobber et al., 2017). Therefore, it might be questionable whether artificial intelligence contains human characteristics, since the human plays the role of a puppeteer that controls the actions of artificial intelligence. Furthermore, the relationship between human agency and machine agency will be important in the era where artificial intelligence is thriving (Sundar 2020). As stated in *Digital Inequalities in the Internet of Things: Differences in Attitudes, Material Access, Skills, and Usage*, "Policies should aim to stress the potential outcomes IoT (Internet of Things) has to offer and should promote transparency and disclosure of how personal data is used as well as better privacy, security practices and regulation" (Deursen et al. 2019).

Conclusion

In conclusion, artificial intelligence is not to be seen as a doomsday scenario, as portrayed in the movies that have been discussed in this article. Yet, it is a means to perceive our own humanity and therefore our own flaws that have to be corrected. Artificial intelligence is a means to confront us with our own intentions. That what makes us trust our fellow human beings is also integral to trust artificial intelligence; transparency. Consequently, by addressing the source of technology, which is humankind, can artificial intelligence be considered as trustworthy, since it is a product of humanity. Nevertheless, it is difficult to assess how much artificial intelligence is operating by its own decision-making or on decisions based on behalf of its creators. Furthermore, Artificial intelligence should not be underestimated. Technology is a powerful entity, as it may cause for destruction or for salvation of humanity. Nevertheless, this is still in the hands of humans who control and decide how to use such technology as artificial intelligence. In this case, the root that perceives artificial intelligence as an evil is humanity itself.

Description of improvement

Thanks to my peers, I had an elaborate feedback that enabled me to make this essay. The first point that was made, was to clearly state my arguments, as to allow for

an insight of the structure of the paper. Consequently, I have produced the following arguments that allowed me to write this essay as such:

Argument 1: Artificial intelligence is to be trusted, since it has been a part of humanity for centuries and therefore an integral part of us as humans.

Argument 2: Artificial intelligence is to be criticized, as it is a developing entity that means it can perform beneficial for humanity or become a threat.

Argument 3: Artificial intelligence is not a doomsday scenario, as portrayed in the movies that have been discussed in this article.

Furthermore, I assumed that my readers understood the concepts that I used. For example, ‘automated decision-making’ has to be more specified as to be acknowledged. Accordingly, I added real-life examples to give the reader an impression on artificial intelligence. Moreover, the tone that I initially set was not in line with the readings and the lecture of the course. For my peers it was too pessimistic and unfavorable towards trusting artificial intelligence. Therefore, I tried to incorporate as much of the lectures and the readings within my writing to portray that artificial intelligence is to be trusted, which in turn also changed my perception on it.

Bibiliography

Araujo, Theo, Natali Helberger, Sanne Kruikemeier, and Claes H. de Vreese. 2020. "In AI We Trust? Perceptions about Automated Decision-Making by Artificial Intelligence." *AI & SOCIETY*, January. https://doi.org/10.1007/s00146-019-00931-w.

Deursen, Alexander J. A. M. van, Alex van der Zeeuw, Pia de Boer, Giedo Jansen, and Thomas van Rompay. 2019. "Digital Inequalities in the Internet of Things: Differences in Attitudes, Material Access, Skills, and Usage." *Information, Communication & Society* 0 (0): 1–19. https://doi.org/10.1080/1369118X.2019.1646777.

Dobber, Tom, Damian Trilling, Natali Helberger, and Claes H. de Vreese. 2017. "Two Crates of Beer and 40 Pizzas: The Adoption of Innovative Political Behavioural Targeting Techniques." *Internet Policy Review* 6 (4). https://policyreview.info/articles/analysis/two-crates-beer-and-40-pizzas-adoption-innovative-political-behavioural-targeting.

Go, Eun, and S. Shyam Sundar. 2019. "Humanizing Chatbots: The Effects of Visual, Identity and Conversational Cues on Humanness Perceptions." *Computers in Human Behavior* 97 (August): 304–16. https://doi.org/10.1016/j.chb.2019.01.020

Hengstler, Monika, Ellen Enkel, and Selina Duelli. 2016. "Applied Artificial Intelligence and Trust—The Case of Autonomous Vehicles and Medical Assistance Devices." *Technological Forecasting and Social Change* 105 (April): 105–20. https://doi.org/10.1016/j.techfore.2015.12.014.

Jonze, Spike. 2014. *Her*. Drama, Romance, Sci-Fi. Annapurna Pictures.

Kubrick, Stanley. 1968. *2001: A Space Odyssey*. Adventure, Sci-Fi. Metro-Goldwyn-Mayer (MGM), Stanley Kubrick Productions.

Leerssen, Paddy, Jef Ausloos, Brahim Zarouali, Natali Helberger, and Claes H. de Vreese. 2019. "Platform Ad Archives: Promises and Pitfalls." *Internet Policy Review* 8 (4). https://policyreview.info/articles/analysis/platform-ad-archives-promises-and-pitfalls.

Marsh, Stephen. 1994. "Trust in Distributed Artificial Intelligence." In *Artificial Social Systems*, edited by Cristiano Castelfranchi and Eric Werner, 94–112. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. https://doi.org/10.1007/3-540-58266-5_6.

McLean, Graeme, and Kofi Osei-Frimpong. 2019. "Hey Alexa … Examine the Variables Influencing the Use of Artificial Intelligent In-Home Voice Assistants." *Computers in Human Behavior* 99 (October): 28–37. https://doi.org/10.1016/j.chb.2019.05.009.

Miller, Tim. 2019. "Explanation in Artificial Intelligence: Insights from the Social Sciences." *Artificial Intelligence* 267 (February): 1–38. https://doi.org/10.1016/j.artint.2018.07.007.

Proyas, Alex. 2004. *I, Robot*. Action, Drama, Sci-Fi, Thriller. Twentieth Century Fox, Mediastream Vierte Film GmbH & Co. Vermarktungs KG, Davis Entertainment.

Sundar, S. Shyam. 2020. "Rise of Machine Agency: A Framework for Studying the Psychology of Human–AI Interaction (HAII)." *Journal of Computer-Mediated Communication* 25 (1): 74–88. https://doi.org/10.1093/jcmc/zmz026.

Turkle, Sherry. 2007. "Authenticity in the Age of Digital Companions." In . https://doi.org/10.1075/is.8.3.11tur.

Winfield, Alan F. T., and Marina Jirotka. 2018. "Ethical Governance Is Essential to Building Trust in Robotics and Artificial Intelligence Systems." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376 (2133): 20180085. https://doi.org/10.1098/rsta.2018.0085.