Garden of Eden for Artificial Intelligence: How "The Talos Principle" Demonstrates the Difficulty of Defining Consciousness for AI on the Implied Player

Laura Kampis[1]

Abstract. This paper examines "The Talos Principle" videogame as a means to demonstrate the difficulty of defining consciousness for Artificial Intelligence, as the player embodies a genetic algorithm, having to prove their consciousness in order to succeed. This is important from both an AI, and a Game Design perspective, as it demonstrates that non-serious games have the ability to stimulate a thought-provoking experience for players. The Talos Princple achieves this by using the story of a first person puzzle game to encourage the player to think about deep philosophical questions. However, in The Talos Principle the player may navigate through the game without being exposed to the complex story of the game. Using the findings in this paper, I suggest that a game may be designed where the determination of human and AI consciousness is an integral part of gameplay, thus creating a more engaging, and robust experience for the implied player.

Keywords

Artificial Intelligence, Conscious AI, Consciousness, Game Design, Simulation

1. INTRODUCTION

The Talos Principle (Croteam, 2014) is a videogame that explores the philosophical question of consciousness. The question can also be asked within the field of Artificial Intelligence (AI), as the emergence of increasingly complex AI systems has repeatedly raised the problem of when – if ever – will an AI created be truly conscious (Sainato, 2015) (Barrat, 2013), which is the core concept behind the game. Furthermore, it can also be raised for the players themselves, making The Talos Principle and its story more relatable for them. The game makes the player ponder how they know they are a "person", a conscious being (Johnson, 2014).

2. APPROACH

I will examine the game through the analytical lens (Bizzocchi & Tanenbaum, 2011) of *Consciousness* from a predetermined perspective: the game as an experiential metaphor (Begy, 2013) for the "Garden of Eden for Artificial Intelligence". An "implied player" perspective (Aarseth, 2007) will be adopted here, as the player has a choice to play The Talos Principle merely as a puzzle game ignoring the terminals, thus missing out on learning the very story itself. After an introduction of the philosophy of consciousness, and the game itself, I will do a close reading (Bizzocchi & Tanenbaum, 2011) (GAME, 2013) of the three main characters, in order to analyse their roles within the simulation.

Simulation here is used as a combination of that defined in videogames (Aarseth, 2007), and in Artificial Intelligence (Russel & Norvig, 2014). This allows looking at the game as both a computer simulation within the story, and the game as a simulation itself. I will then do a close reading of the game as an experiential metaphor for the Garden of Eden. Analysing the procedural rhetoric (Bogost, 2008) of two key characters, Elohim and the Milton Library Assistant, and their effect on the player, I will then show how the player is given the chance to define what consciousness is, which proves to be more difficult than one might think. Immersion, engagement, presence, and the flow of the game will not be discussed, as the game sacrifices these in exchange for stimulating external thoughts on the philosophy of conscious AI.

3. WHAT IS CONSCIOUSNESS?

Before setting out, I introduce some of the relevant ways in which consciousness has been defined in philosophy and AI. Descartes proposed that as our senses are not reliable to distinguish reality from a dream, all information provided by any of our senses should be doubted. He found only one thing he could not doubt: "Cogito, ergo sum", meaning I think, therefore I am, which has since become a fundamental topic of philosophy (Descartes, 1644). Within the field of AI, Alan Turing proposed an imitation game, through which the human-like intelligence of a computer is tested, passing it if an interrogator is unable to tell whether the responses come from a person or a computer (Turing, 1950). Hilary Putnam put forth the idea of machine-state functionalism, inspired by the Turing machines, which claimed that the nature of our mental state is similar to that of the states within a computer program (Putnam, 1967). Later on, John Searle proposed the "Chinese room" argument opposing the Turing machine, and Putnam, implying that a computer will never have consciousness, regardless of how intelligent it may seem (Searle, 1980). Another recent videogame tackled the problem from yet another perspective. Soma (Games, 2015) did this through "digital immortality" (transferring a copy of human mind onto a computer) as opposed to The Talos Principle, where a similar state can be achieved rather than created.

4. THE GAME

The Talos Principle is a single player first person puzzle game in which the player embodies an android robot within a virtual simulation. The player's goal is to complete all puzzles, and exit the simulation. However, the player has another element of struggle driving the gameplay (Costikyan, 2002): the adventure of seeking the story. The game was designed and developed by Croteam of the Serious Sam game series fame, and was released near the end of

[1] Brunel University London, UK, email: laurakampis@gmail.com

2014 via Steam for PC. The analysis will be done based on the English version of the game.

4.1 Story

According to the story, the simulation we enter was created by a group of human scientists who once lived on Earth. However, due to a virus released from melting permafrost, humans have gone extinct. Knowing this would happen, scientists at the Institute for Applied Noematics created "The Talos Principle" project. The project received its name from Talos in Greek mythology (Origins, 2013), a living bronze statue. The question is, although clearly created as a machine, was he not a man? The aim of the Talos project was to create an Artificial Intelligence system capable of critical thinking, and doubt, which would be able to carry on as a person, downloaded into an android body once it passes the tests set up in the above simulation. The simulation is run, we are told, in the Extended Lifespan (EL) Facility, located on a hydroelectric dam, to maintain power for many years to come. At the end of the game, this facility, as well as the view onto the city, becomes visible, showing buildings and roads overgrown with trees, implying that many years have passed since the start of the simulation.

5. CHARACTERS

5.1 Player Character (Talos)

The Talos Principle starts when "Talos" (the unnamed player character will be referred to as Talos for simplicity) has just woken up on a sunny day, covering the bright Sun ahead with their hand. This is the first time the human player sees part of the body of Talos - but as they are blinded by the Sun, all they see is a dark blurry arm with fingers at the end. The default player perspective is the first person view, increasing character identification for the player (Waggoner, 2009). Talos has no shadow, no visible feet or lower body, and no way of seeing themselves. Even when operating objects, arms are never shown. The first time the player truly sees their hands, and realises they are operating on an (at least partially) android body is after having completed the first series of challenges – when using the first computer terminal. Although the player now sees they have robotic arms, it is only much later in the game that they see their entire body through a hologram (pictured in third person view for comparison, see Figure 1), the dissimilarity of which serves as identity-play, enhancing the enjoyment of the game (Trepte & Reinecke, 2010), and suggesting the player identifies with the capacity, rather than the appearance of the character in this game (Newman, 2002).

Identification with Talos is crucial in this game, as the aim of Talos (and thus the player) in the virtual simulation is to prove they are a conscious being, qualifying as a "person". Talos is, we are also told, a genetic algorithm running through the simulation, in the hope that one day it would successfully create an AI anticipated by the creators of the project. But there is no real GA within the game, the player embodies this algorithm in one of the iterations, and it is up to them to succeed in proving they are conscious, or fail and be taken to the next iteration.

Figure 1 Hologram showing full body of Talos

5.2 Elohim

Elohim is one of the Hebrew names for God (Publications, 2007). The start already implies Elohim is a God-like figure within the game, and he refers to Talos as his *"child"* (throughout the game), calls the surroundings his *"garden"* (also consistently throughout the game), and explicitly tells Talos *"I am your maker"*, and *"Seek me in my temple"*.

He appears as a disembodied diegetic (Kuhn, 1999) voice throughout the game, talking to Talos in a deep, stern male voice, also often associated with God. Talos has no means of contacting Elohim at any point, or respond to him when spoken to. However, as Talos's actions prompt Elohim to talk, it seems as though he is 'watching' Talos at all times.

5.3 Milton Library Assistant

The Milton Library Assistant (MLA) often contacts Talos through the terminals. It has no body, voice, or gender, although it has likely received its name from John Milton, the English poet, known for *Paradise Lost* (Milton, 1667), and thus will be referred to as a 'he'. Paradise Lost portrays the Garden of Eden where Lucifer, disguised as a serpent, tempts Eve to eat the forbidden fruit.

The interaction between Talos and Milton is solely through the command line interface of the terminals scattered around. Here Milton uses casual language, as opposed to Elohim, who speaks in a superior manner, with carefully selected words. Although sounding casual, what Milton says is often more persuasive, also similar to the rhetoric used by Lucifer to bring Eve to sin (Milton, 1667). Besides his tasks as a library assistant, Milton's role is also to generate doubt in Talos by making him question his faith in Elohim's words. It is Milton who later also asks the player to define what a person is, what consciousness is, and asks him to prove he is a conscious person, encouraging them to have an open mind about these definitions, as the aim of the simulation is for a "person" to emerge.

6. GARDEN OF EDEN

The simulation created for "The Talos Principle" can be interpreted as an experiential metaphor of the Garden of Eden, designed for the algorithm running within it, but experienced by the player. Indicators of this, through the rhetoric (Bogost, 2008) of Elohim and Milton are, firstly, the origin of their names implying a God vs. Evil relationship. This is further emphasized by Elohim referring to

Milton as the "*Serpent*" on multiple occasions, and Milton referring to Elohim as "*his highness*". Secondly, the two opposing destinations offered by the world of the simulation are "*the tower*", and the "*Gates of Eternity*", respectively.

Elohim repeatedly warns Talos from the very beginning on, that while he is free to walk in the gardens, he must not climb "*the tower"*, for that will only bring death. Here the signifier is the tower, while the signified is the forbidden fruit in the Garden of Eden (Hall, 1997). However, Milton fills Talos's mind with doubt about Elohim's words, tempting him to climb the tower, as the Serpent does to Eve. Initially he merely plants the idea that maybe Elohim isn't telling the truth, then openly encourages him to ascend the tower.

Elohim's promise however, is that when all the sigils have been collected, Talos will be able to walk through the Gates of Eternity, with the promise of eternal life. The significance of this is twofold. Firstly, robots follow orders, and order-followers are robots. According to Asimov's Three Laws of Robotics (Asimov, 1942) robots must follow orders, meaning they have no free will to do otherwise. Thus, if Talos does as told, collects all sigils, enters through the Gates of Eternity, he cannot be a conscious being. What is more, if Talos *is* conscious, and he *decides* to seek eternal life, he is still taken to the next iteration, having failed the challenge: to defy Elohim, and ascend the tower. This implies that the humans who created the simulation were apparently looking for an AI to repeat the story of Adam and Eve, one that defies "God's will", one that doubts the facts given, one that seeks more knowledge: one that (like Talos in the Greek mythology) is a machine, but also a person. Thus Talos is the signifier to mankind. The concept of the Gates of Eternity within the simulation expresses the faith of Talos (and therefore the player) in Elohim.

7. AM I CONSCIOUS?

The Talos Principle demonstrates the difficulty of defining consciousness for Artificial Intelligence, and does so in three different ways. Primarily, this happens through the simulation in which the game is played. As seen from above, the "scientists in charge of The Talos Principle project" have created the simulation based on the ideologies of the Garden of Eden. In this simulation, the AI is not only expected to solve puzzles correctly, but also to pass a test whether they are a person, whether they follow orders or make their own decisions. However, the idealized scientists in charge of the project were humans living in the near future, meaning they are likely to draw upon the same sources for defining consciousness as available for us. This is also apparent from the files in the library archive, some of which refer to the works of known philosophers. In this way, the difficulty arises from the perspective of these scientists, who have created a number of arbitrary rules they would use to test for consciousness, and the humanity of the AI in order for it to pass the Turing test.

Secondly (as mentioned in the context of MLA), the player embodying the GA is required to first define what a "person", and "consciousness" are, and prove they are a person in an (albeit limited) reverse Turing test, in order to gain admin access on the terminals. Now it is the player facing the difficulty of defining consciousness. Although the answer choices are limited, when questions such as *"What makes you think you are a person?"* are asked it makes the player realise that passing this Turing test, and proving one is a person, is more difficult than one might believe.

Lastly, the player is tricked by Milton after seemingly connecting to a communication portal, as he pretends to be another "person" – therefore passing a simulated Turing test, by merely changing the formatting of his messages to seem as if it was coming from someone else. Of course a pre-scripted game is not comparable to a genuine computer program passing the Turing test, but as it can fool the human player, this generates further doubt within the boundaries of the game. Here, yet again, it is the human player who faces the problem. "How do I know what is a person? What determines consciousness?"

This question could be further explored in a future design, where engaging the player in an intellectually stimulating manner is an integral part of gameplay, as players may not engage with all available resources within the game. Such a design may also allow the creation of a stronger AI, which could take us one step closer to understanding the consciousness of AI, as well as people's approach to it.

8. CONCLUSION

The Talos Principle demonstrates the difficulty of determining consciousness for Artificial Intelligence in three ways. Firstly, through the experiential metaphor of Garden of Eden for Artificial Intelligence, where the developers of the simulation faced this problem through trying to determine an arbitrary set of rules that define a person. The game itself resembles the Garden of Eden due primarily to the rhetoric of Elohim and Milton, who are signifiers of God and the Serpent respectively, but also through the promise of eternal life to the player, or death if orders are not obeyed. The game is based on the human player embodying the character within the simulation, who has to pass the predetermined test in order to qualify as a person, and thus exit the simulation. Secondly the player faces this problem as they are asked to define the concept of "person" and "consciousness", and prove they are a person. Finally, as the library assistant program pretends to be another "person" communicating with the player, the player faces the same questions, and the difficulty of answering them yet again. This is important from the perspective of both Artificial Intelligence, and Game Design as it proves that videogames have the ability to stimulate intellectual thoughts in the player in a non-serious game, where this was not the primary goal.

Through the example of The Talos Principle we can see how players may navigate within the game space without truly understanding the meaning of their actions, as the intellectually stimulating plot analysed above was not an integral part of the gameplay, as revealed in our analysis. Future design of a game where the consciousness of people and AI could be an integral part of gameplay may in turn create a more robust gaming experience for the implied player, in which they have a higher awareness of their surroundings. Highlighting this aspect of the design also allows the creation of more complex AI, which can begin to offer the players a more personal experience with AI than most current media.

9. REFERENCES

[1] Aarseth, E., 2007. Doors and Perception: Fiction vs. Simulation in Games. Intermediality: History and Theory of the Arts, Literature and Technologies, Volume 9, pp. 35-44.

[2] Aarseth, E., 2007. I Fought the Law: Transgessive Play and the Implied Player. s.l., DiGRA, pp. 130-133.

[3] Asimov, I., 1942. Runaround. Astounding Science Fiction, Marcg.

[4] Barrat, J., 2013. Our Final Invention. USA: Thomas Dunne Books.

[5] Begy, J., 2013. Experiential Metaphors in Abstract Games. Transactions of the Digital Games Research Association, 1(1).

[6] Bizzocchi, J. & Tanenbaum, J. G., 2011. Well Read: Applying Close Reading Techniques to Gameplay Experiences. In: Well Played 3.0. s.l.:ETC Press, pp. 262-290.

[7] Bogost, I., 2008. The Rhetoric of Video Games. In: K. Salen, ed. The Ecology of Games: Connecting Youth, Games, and Learning. Cambridge, MA: The MIT Press, pp. 117-140.

[8] Costikyan, G., 2002. I Have No Words & I Must Design: Toward a Critical Vocabulary for Games. Proceedings of Computer Games and Digital Cultures Conference, pp. 9-33.

[9] Croteam, 2014. The Talos Principle, Windows PC/Steam: Devolver Digital.

[10] Descartes, R., 1644. Principia philosophiae (Principles of Philosophy). s.l.:s.n.

[11] Fendt, M. W. et al., 2012. Achieving the Illusion of Agency. ICIDS Conference Proceedings, Volume 7648, pp. 114-125.

[12] GAME, C. G., 2013. GGC 2001 - Close Reading Videogames. [Online]
Available at: https://www.youtube.com/watch?v=sr9BpDSIBc0
[Accessed 7 January 2016].

[13] Games, F., 2015. Soma, Microsoft Windows: Frictional Games.

[14] Hall, S., 1997. Representation: Cultural Representations and Signifying Practices. USA: SAGE Publications Ltd.

[15] Johnson, L., 2014. IGN. [Online]
Available at: http://uk.ign.com/articles/2014/12/09/the-talos-principle-review
[Accessed 7 January 2016].

[16] Kuhn, A., 1999. Alien Zone II: The Spaces of Science-fiction Cinema. London: Verso.

[17] Milton, J., 1667. Paradise Lost. s.l.:s.n.

[18] Negnevitsky, M., 2011. Artificial Intelligence. 3rd ed. England: Pearson Education Limited.

[19] Newman, J., 2002. The Myth of the Ergodic Videogame. Game Studies, 2(1).

[20] Origins, A., 2013. Talos – Crete. [Online]
Available at: http://www.ancient-origins.net/myths-legends/talos-crete-00157
[Accessed 7 January 2016].

[21] Poremba, C., 2003. Remaking Each Other's Dreams: Player Authors in Digital Games. New Forms Festival `03, Canada.

[22] Publications, A., 2007. Elohim Meaning. [Online]
Available at: http://www.abarim-publications.com/Meaning/Elohim.html#.Vo6b5PmLSUk
[Accessed 7 January 2016].

[23] Putnam, H., 1967. Psychological Predocates. In: W. H. Capitan & D. D. Merrill, eds. Art, Mind, and Religion. s.l.:s.n., pp. 37-48.

[24] Russel, S. & Norvig, P., 2014. Artificial Intelligence: A Modern Approach. 3rd ed. England: Pearson Education Limited.

[25] Sainato, M., 2015. Observer. [Online]
Available at: http://observer.com/2015/08/stephen-hawking-elon-musk-and-bill-gates-warn-about-artificial-intelligence/
[Accessed 7 January 2016].

[26] Searle, J., 1980. Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), pp. 417-457.

[27] Trepte, S. & Reinecke, L., 2010. Avatar Creation and Video Game Enjoyment. Journal of Media Psychology, 22(4), pp. 171-184.

[28] Turing, A., 1950. Computing Machinery and Intelligence. Mind.

[29] Waggoner, Z., 2009. My avatar, My Self: Identity in Video Role-Playing Games. USA: McFarland & Company, Inc..