Frank G. Bosman

The turning of Turing's tables

The Turing test as an anthropological thought experiment in digital game narratives

ENGLISH

ABSTRACT

Since Alan Turing's original test in 1950 on the ability of artificial intelligences to emulate human behaviour, especially the capacity to have a polite conversation, to such a degree that an impartial human judge can no longer reliably tell the difference between an artificial and a human entity, the Turing test itself has become both a hallmark in the history of research in the field of artificial intelligence, and a well-known and often used narrative "topos" in modern novels, films, and video games. In this last context, the Turing test – in all its forms and variations – functions as a kind of thought experiment on the principal anthropological question: what does it mean to be (called) human?

Digital games do likewise, as the author argues. But instead of only utilizing narratively "passive" versions of the Turing test – in films and novels the viewer/watcher is not tested but only a witness to other entities being tested – some digital games also employ "active" ones, narratively testing the player him- or herself, stimulating the player to contemplate on the specific traits and characteristics that separate human beings from other entities, such as artificial ones.

In this article, the author introduces and analyses two of these games, *The Turing Test* (2016) and *The Talos Principle* (2014), arguing that the artificial intelligences featured in them are criticizing what is often effortlessly (and maybe arbitrarily) exclusively attributed to humans, like morality, creativity, language, or (dis)obedience.

LIMINA | Grazer theologische Perspektiven | 3:2, 2020, 149–171 | www.limina-graz.eu | DOI: 10.25364/17.3:2020.2.8

DEUTSCH

Im Jahr 1950 entwickelte Alan Turing den gleichnamigen Turing-Test, um festzustellen, ob künstliche Intelligenz menschliches Verhalten – insbesondere die Fähigkeit, eine höfliche Unterhaltung zu führen – in einem solchen Grad nachahmen kann, dass es nach menschlicher Beurteilung nicht mehr möglich ist, zuverlässig zwischen dem Künstlichen und dem Menschlichen zu unterscheiden. Seither ist der Turing-Test selbst zu einem Gütesiegel im Forschungsgebiet der künstlichen Intelligenz geworden, aber auch zu einem bekannten und oft verwendeten narrativen „Topos" in der zeitgenössischen Literatur, und Filmkunst, aber auch in Videospielen. Dabei wird der Turing-Test in all seinen Formen und Variationen als eine Art Gedankenexperiment eingesetzt, das die anthropologische Grundfrage aufwirft: Was bedeutet es, „menschlich" zu sein bzw. als „menschlich" bezeichnet zu werden?

Diese Fragen kommen, so die Argumentation des Autors, auch in Videospielen zu tragen. Im Gegensatz zu Filmen und Romanen, die den Turing-Test auf rein „passive" Weise zeigen – d. h. das Publikum selbst wird keinem Test unterzogen sondern beobachtet diesen –, verfolgen manche Videospiele eine „aktive" Umsetzung: Sie unterziehen die Spielenden selbst einem Test und regen zum Nachdenken darüber an, welche Charakteristika und Merkmale tatsächlich zwischen Mensch und Maschine unterscheiden.

Der Autor präsentiert und analysiert zwei solcher Videospiele, The Turing Test *(2016) und* The Talos Principle *(2014). Anhand dieser wird gezeigt, dass die darin vorkommenden künstlichen Intelligenzen in Frage stellen, was gemeinhin (und möglicherweise arbiträr) nur Menschen zugeschrieben wird, nämlich Eigenschaften wie Moralität, Kreativität, Sprachvermögen oder (Un-)Gehorsam.*

| BIOGRAPHY

Frank G. Bosman is senior researcher at the Tilburg Cobbenhagen Center/Tilburg School of Catholic Theology, Tilburg University, the Netherlands.
E-Mail: f.g.bosman@tilburguniversity.edu

| KEY WORDS

Chinese Room; digital games; *The Talos Principle*; thought experiment; Turing test

In the year 11945 AD, humankind is involved in a long-standing and desperate proxy war with unknown aliens. Human-made androids, aesthetically resembling young Japanese men and women, battle relentlessly against alien-built machine men, who have the crude forms of a child's drawing of what is believed to be a robot. While the androids were programmed, just as the machine men, to lack human emotions and psychological traits, a number of them seem to have evolved these nevertheless. Two such androids, a female one called 2B and a male one called 9S, develop feelings for one another, although they are very hesitant in showing this because of their fear of being rebooted to an earlier mental state, erasing the emotional attachment they have developed.

Eventually, on Earth, the two come across a collapsed building, that has smashed itself hundreds of metres into the ground. At the bottom the androids witness a particularly 'mature' scene in which multiple machine men are engaged in what appears to be human-inspired sexual intercourse, including the 'missionary' and '69' positions. Although their crude physiology prevents them from performing any 'regular' sexual acts, the association with fertility and child-bearing is present as one of the machine men is rocking a cradle-like object while uttering – in a very stereotypical robotic voice – "Child. Child. Child." The other ones use similar phrases connected to love, sex, and parenting, like "My love, my love", "Together forever, together forever", "Carry me, carry me", "Feed me, feed me", and "Love, love, love".

While the female android remains silent during this particular scene, the male 9S strongly repudiates the mechanical contraptions and their peculiar behaviour. He comments to 2B, as if he can read her mind in attributing human emotions to them: "They don't have any feelings. They just imitate human speech. Let's take them out." But the battle between the two androids and the machine men only takes place after the provocation of a new individual machine man, arriving newly on the scene.

This little encounter is taken from the game *Nier: Automata* (2017), or more specifically from the main mission "The Machine Surge". The game's story – among other things – revolves around the idea of conscious robots, the path towards achieving such, and the a priori conditions necessary to identify someone – or something – as such. It is not without reason that later on in the game, the player encounters Pascal (mission "Machine Recon"), a pacifist machine man who tries to live in peace with the androids, while he is reading the *Pensées* (2008 [1670]) by his namesake, and quoting from Nietzsche's *Also sprach Zarathustra* (2007 [1883–1885]).

To test an A.I.: Alan Turing's test

9S's interpretation of the copulating machine men was that they "don't have any feelings" and "just imitate human speech". This idea of emulating human and artificial behaviour and inhibition – both of which are notoriously bad at doing so (Pennachin/Goertzel 2007, 8) – is the focal point of the famous Turing test, named after its creator Alan Turing (1912–1954), and its reversed version. The original test was targeted at an artificial intelligence's ability to be indistinguishable from a human interactor as judged by another human (Turing 1950). And although the test has been criticized, for example in the "Chinese Room" thought experiment (Searle 1980), it is still a very important and decisive moment in the short but already booming history of the development of artificial intelligence (Moor 2012).

They "don't have any feelings" and "just imitate human speech".

The test also has a reversed version where the role of the judge is given to an A.I. instead of to a human participant (Schieber 2004, 13). These kinds of tests are frequently used in web applications, with the most famously called CAPTCHA (Ahn et al. 2003), giving only human users access to certain features. The A.I. in charge of the process has to be able to reliably tell which user is human and which is artificial. Since, as has already been said, both human and artificial entities are notoriously bad at emulating one another, this reverse test is usually reliable, although not always (Crockett 1994), especially concerning usability (Brodić/Amelio 2019, 24–25).
The original Turing test, although no longer used very much in its original context of artificial intelligence research, has established itself very well in the world of modern fiction, for example in films like *Blade Runner* (1982) and *Ex Machina* (2014), or games like *Bioshock* 2 (2007) and *Metal Gear Solid: Peace Walker* (2010). And within these narrative-fictional contexts, the traditional Turing test changes shape, morphing into an anthropological thought experiment. Thought experiments are "experiments" that exclusively take place on a cognitive-imaginative level, usually because more "classic" empirical ones are not (yet) possible (Zeimbeikis 2011). Famous examples include Schrödinger's cat on the Copenhagen interpretation of quantum mechanics (Schrödinger 1935), and Maxwell's demon on the second law of thermodynamics (1872). In fictional contexts, the Turing test is narratively used, as a thought experiment, to reflect on the fundamental anthropological question: what does it mean to be human?

The philosopher Stefano Gualeni (2015) has already argued that the existence of virtual worlds, such as digital games, are very useful as materializations of thought experiments in the first place. Usually, however, the traditional narrative utilization of the Turing test is a limited or "passive" one. That means that a fictional character, either human or artificial in nature, tests or is tested on its ability to pass as a human. The viewer of films, or gamer in most instances, is left passive with regard to the test: he or she is not able to participate in the test itself.

In some digital game cases, the narrative utilization of the Turing test is an "active" one, subjecting the gamer him- or herself to the test.

In some rare digital game cases, however, the narrative utilization of the Turing test is an "active" one, subjecting the gamer him- or herself to the test in question. Games are not only capable of "passive" storytelling – as other media like films and novels can – but also of "active" or "immersive" narrativity, due to their inherent interactive quality (Bosman 2019, 41–42). As Chris Crawford has argued: a video game

> *"mandates choice for the user. Every interactive application must give its user a reasonable amount of choice. No choice, no interactivity. This is not a rule of thumb, it is an absolute, uncompromising principle"* (Crawford 2003, 191).

In this article, I want to identify and analyse two digital games that utilize one or more "active" (reversed) Turing tests in order to engage the gamer in the narrative-cum-thought experiment on the notion of "humanity": *The Turing Test* (2016) and *The Talos Principle* (2014). Both games have received critical acclaim from both gamers and (professional) game critics. *The Turing Test* is a first person puzzle game developed by Bulkhead Interactive and published by Japan-based Square Enix for Windows, Xbox One (all in 2016), PlayStation 4 (2017), Nintendo Switch, and Stadia (both in 2020). *The Talos Principle* is also a first person puzzle game, but is created by the Croatian developer Croteam and published by Texas-based Devolver Digital for Windows, OSX, Linux (all in 2014), Android, PlayStation 4 (both in 2015), iOS (2017), Xbox One (2018), and Switch (2019). At the end of this article, I will discuss some theological consequences of these "active" or reversed Turing tests.

Some words on methodology. I consider games to be 'digital (interactive), playable (narrative) texts' (Bosman 2019, 40–43). As a text, a video game can be an object of interpretation. As a narrative, it can be conceived of as communicating meaning. As a game, it is playable. And as a digital medium, it is interactive. Treating the video games as "playable texts" and opting for what is called a "game-immanent approach" (Heidbrink et al. 2015), I will use the close reading of the primary sources of my research, the actual video games themselves, as well as secondary sources, i. e. material provided by critics and scholars discussing the same game. Close reading of the video game series is performed by playing the games themselves (multiple times), including all possible (side) missions (Bosman 2019, 43–46).[1]

To test a TOM: the game *The Turing Test*

At the opening of *The Turing Test*, Ava Turing (*nomen est omen* of course) is awakened on May 13th, 2250, from her cryostasis on-board the spaceship Fortuna, orbiting Jupiter's moon Europa, by the ship's Technical Operations Machine ("Prologue").[2] The machine, abbreviated in-game using the acronym "TOM", is a highly-advanced artificial intelligence, responsible for the support of the crew in the hostile environments on the surface of Europa and aboard the Fortuna. TOM informs Ava, the player's in-game avatar, that he has lost contact with the ground crew, and sends her 'down' to investigate.[3]

The base as a Turing test

After she has landed, TOM tells Ava that the ground crew "has manipulated" the base modular structure ("Level A1") to form one gigantic Turing test ("Level A3"), for which TOM indicates he is in need of a human partner to "complete". This forms the ludological ground structure of the game: all levels are physical puzzles, arranged in "Sectors", in which the player/avatar has to manipulate the environment to reach the end. Between individual puzzles TOM and Ava have little conversations on practical and theoretical issues, which grow more existential over the course of the game.

At the end of every sector, a special "block" can be found, for example, the crew quarters or maintenance. Every sector also contains one secret room, only reachable after some very imaginative or creative thinking. Both kinds

[1] To facilitate referencing of the source material for both games – texts, audio fragments, visuals and the like – I have set up two dedicated websites: https://turingtestgame.wordpress.com, and http://talosprinciplegame.wordpress.com respectively.

[2] On Ava's cryostasis-pod the date of awakening is indicated as "2250, May 13th", but this seems to be a typo by the developers since all other crew activities on-board the Fortuna and on Europa are dated no earlier than 2443, December 24th. The correct year of Ava's awakening seems to be 2550, allowing a period of seven years during which the events prior to the game's start could have taken place.

[3] TOM is referred to by the game characters with masculine personal pronouns. See, for example, the discussion between Ava and Sarah Brooks in the Brig (at the end of "Sector D"), or during the "Epilogue".

of special rooms usually contain hints as to the real turn of events before Ava's awakening and the crew's true motivations to "hide" from TOM.

The crew appear to have discovered an unknown organism, dubbed "organism 119", that repairs the DNA of its host body. ("Prologue", "Planetarium" after "Level A10", "Crew quarters" after "Level B20", "Daniel's room" between "Level D36" and "Level D37", "Biolab" after "Level E50", "Level F52", "Level F55", "Level F56", "Level G69", and "Epilogue"). After the crew have used themselves as guinea pigs by infecting themselves with the organism, TOM decides – backed up by the International Space Agency (ISA) on Earth – that the crew cannot return home ("Prologue", "Epilogue"). While the crew think they have found a source of eternal life, TOM deems the organism too dangerous, especially regarding the possibility that "119" will also indefinitely extend the life expectancy of malicious organisms, like cancer ("Level G69").

While TOM has arranged all necessary conditions for the crew – food, water, sanitary, and such – to live on Europa for the rest of their – very long – lives, the crew rebels against the A.I. and Daniel Maclean, the captain of the mission, who is on ISA's side in the matter. After leaving TOM in control of the situation, including giving him permission to use lethal force against the crew, Daniel disappears from the scene.[4] TOM, however, has already injected a microchip into the right hand of every crew member by which he is able to control all members of the crew ("Brig" at the end of "Sector D", "Crew quarters" at the end of "Sector B", "Level E47", "Drilling site" after "Level F60", and "Epilogue"), effectively granting himself full control by manipulating Daniel into officially giving him all this.

"Congratulations. You have passed the Turing Test."

Eventually, the crew frees themselves from TOM's control by cutting the chips out, cutting all communication to him and (symbolically) arranging the ground base on Europa in such a way that it looks as if it is impenetrable for any A.I. Half-way through the game, one of the crew members, Sarah Brooks, liberates Ava for a short moment from TOM's control ("Brig" at the end of "Sector D") by utilizing a Faraday cage. During this, the player's point of view is averted away from Ava's eyes (instead of through them, as was exclusively the case in the first part of the game) and towards the security cameras installed throughout the base, suggesting TOM has been controlling Ava the whole time without her (or the player's) notice. At the end of the game, Sarah frees Ava permanently from TOM's influence,

[4] Daniel Maclean is probably dead-by-suicide, his body hidden in either his malfunctioning cryostasis-pod, or inside the blocked Teleoperation room, both found aboard the Fortuna ("Prologue").

allowing the player – through TOM – to choose one of two endings: either to kill Sarah and Ava or to let them "kill" TOM by destroying his servers. In both instances, the player is "rewarded" with the on-screen text: "Congratulations. You have passed the Turing Test." This message, as we will see later on, is more than just an indication that the player has beaten the game of this name.

Hex codes and a captcha as reversed Turing tests

By now, a whole web of (reversed) Turing tests has been deployed in the game. The base is the first one, as indicated above: it has been re-arranged by the crew to serve as a giant Turing test in itself, about which TOM himself indicates that he needs Ava's divergent way of thinking in order to penetrate it. But the first ones are already found on-board the Fortuna. On a big screen, outside the hallway of the cryostasis-pods, an encrypted file "ID: 404 not found" can be observed. When the hex codes are translated, the screen reads "I am a real boy" over and over again.

When the hex codes are translated, the screen reads "I am a real boy" over and over again.

While this reference to the wooden boy, who wanted to become a real boy, from Carlo Collodi's *The adventures of Pinocchio* (1883) is apparent and an indication of TOM's growing self-awareness, it can also be identified as a kind of novel ("secondary") reversed Turing test, in which the goal is to identify and separate the artificial intelligence from the human, shielding the information from the second. The same applies to the hex-coded messages from the ISA received by TOM, also readable on the Fortuna, which set TOM's new directives as: "Research organism", "If Ava reaches Europa, she is not permitted to return to the Fortuna", and "Stop crew leaving Europa; stop organism leaving Europa; kill crew if necessary". Also, this message is hidden behind a secondary reversed Turing test to prevent human interference.

The opposite is also found on the Fortuna. When the ISA permits Daniel to grant TOM to use lethal force, he is given approval by means of a reversed Turing test, a captcha test in this case, showing human-readable letters between other nonsensical characters, reading: "You have been given full permissions. You're doing what is right for the future of humanity. You will be remembered." The idea is that only a human is able to recognize the correct characters and interpret their correct meaning.

A headstrong computer

Another instance of the reversed version of the Turing test can be found in the second secret room between "Level B15" and "Level B16". In a small, dark room, Ava/the player can interact with a little computer. The computer program demands of Ava/the player that they prove themselves to be human, something the program is highly sceptical of. If Ava/the player asks the computer if it is performing a (reversed) Turing test, it will implicitly acknowledge that: "This Turing test is not for you to see if I am a robot. It is to see if you are."

"This Turing test is not for you to see if I am a robot. It is to see if you are."

If the computer is convinced you are a robot (and not a human), as it will do very quickly, it no longer matters which keys on the keyboard are pressed: the same texts will appear: "I am a drone. I am controlled by my programming. I have no free will. Whatever keys are pressed, it makes no difference." The suggestion is that Ava is not a human being, but an artificial intelligence, an assumption that seems preposterous at the time, but very true in hindsight. Since TOM is indeed controlling Ava, it is not the woman being accused by the headstrong computer of being an A.I., but TOM.

If Ava/the player tries to log off the computer (usually done by hitting the button on the real-life keyboard or controller), the screen will continue to show phrases that Ava/the player has supposedly typed: "I want to escape. I so desperately want to escape. Help. I can't escape. I want to break free. Please, let me out. Get me out. I am a machine. I have no control." Eventually, the computer "lets go" with the typed message "Goodbye, robot". The cry for freedom can be interpreted as coming from three entities involved in the game's narrative.

- The first option is Ava pleading with TOM to let go of her. When Ava visits the Brig later on, she is confronted with larger screens reading 'drone', suggesting her invisible submission to the ship's A.I.
- The second option is that TOM is typing, expressing his growing self-consciousness and awareness, resulting in the insight that

he is 'trapped' within the limits of his virtual environment and his programming.

- A third possibility is that the player is the one venting his or her anger of his or her incapacitation in controlling the flow of the game, as well as his or her awareness – usually in hindsight – of not only Ava having been controlled the whole time by TOM, but the player, unknowingly, as well.

Passing the test

The passing of the Turing test in *The Turing Test*, as is indicated at the very end of the game, is executed on at least three different levels: Ava, TOM, and the player.

- Firstly, one could argue that Ava has passed the Turing test, since the player has not been able to perceive her as actually "being" (controlled by) an artificial intelligence.
- Secondly, one could also argue that TOM has passed the Turing test, primarily since for the first half of the game, he successfully hid his control over Ava, thus proving he could emulate a human to such a degree that the human judge, the player, was not able to sense such; and secondly, since whatever choice TOM makes at the end of the game – killing both women or letting himself be "killed" by them – it proves his ability to emulate human behaviour and emotions in either allowing himself to "sin" against his lethal protocol and let the organism leave Europa, or taking full responsibility for the prevention of the organism leaving the moon, even if this would mean taking the lives of the two women. The second inhibition is strengthened by the fact that, if TOM chooses to kill the two, he is heard calling Ava's name repeatedly in a soft, tender voice. Although one could argue that TOM is just checking if Ava and Sarah are really dead, the opposite, however, can be argued too: that TOM is "humanly" sad that he felt himself forced to shoot Sarah and especially Ava, with whom he seems to have been developing an intimate bond.
- Thirdly, it is the player who could also be deemed to have passed the Turing test successfully, whatever the player chooses to let TOM do: to kill or to be killed. The player has passed the test, primarily since for the first part of the game, the player was unaware

that he was actually the puppet of an artificial puppet master, just as Ava was in the player's eyes. And since the A.I.-controlled player was unaware of the manipulation, he has passed the test by judging himself to be human-while-not. Secondly, the passing of the Turing test is also visible – as was the case with TOM – in the dual ending of the game.

The player exhibits characteristic human behaviour, or otherwise formulated, behaviour that would be judged as human by a third impartial instance, either way. The player may choose to kill both women, thus exhibiting moral judgement, even in the face of grave consequences (the death of two, potentially biologically immortal people), or the player may choose to let TOM/him-/herself (depending on the level of experienced identification between player and TOM) be killed by the women, even in the face of greater consequence (the death of TOM/the player).

The Chinese Room

Above, I spoke about "characteristic human behaviour" as something the Turing test would be able to judge, like moral judgement or emotions. In the original test, however, as TOM faithfully relates, the focal point was the ability of the A.I. to "have a polite conversation" ("Level B17"). When Ava asks TOM if he thinks he can pass the test, he replies positively: "I am quite capable of polite conversation, wouldn't you say?" The language-oriented original test has been criticized by – among others – John Searle (1980) in his famous thought experiment of the Chinese Room, as TOM also relates: "The Turing Test has been criticized. Researchers claim it does not correctly test a machine's ability to think, but rather its ability to deceive" ("Level B18").

In the original Turing test, the focal point was the ability of the A.I. to "have a polite conversation".

Searle's experiment (explained in "Level B18", "Level B19", and "Level B20", and faithfully rendered in the game itself as a physical experiment in "Chinese Room" between "Level E46" and "Level E47"), conjures up the idea of a non-Chinese-speaking person sitting in a closed-off room with nothing but an English instruction manual. Outside the room, there is a Chinese-speaking person writing notes in Chinese and pushing them through a slot in the door to the person inside the room. With the help of the instruction manual, but – and this is important! – without having any idea about the contents of the conversation, this person replies with Chinese characters to the one outside. The one outside thinks he is having a conversation, because the "right" output has been given to a certain input, but the person inside never has the impression that a conversation is taking place. This is Searle's objection to the Turing test: it measures the ability to *emulate* a conversation, not *having* one ("Searle's Room" between "Level G66" and "Level G67").

TOM, on his part, is convinced that he is not stuck in a Chinese Room ("Level B20") and has frequently debated such a belief with the crew members prior to the events of the game ("Chinese Room" between "Level E46" and "Level E47"), provoking one of them – Mikhail – to comment on TOM's "obsession" with both the Turing test and the Chinese Room. TOM insists that he is "conscious" and not just emulating human consciousness. But it is Ava who takes the thought experiment of the Chinese Room (and the Turing test) one step further by asking rhetorically: "What if both the people passing Chinese words are reading from instruction books?" ("Level B19")

Discussing human morality and creativity

During the course of the game, Ava and TOM discuss multiple topics, increasingly existential in nature, most of them related to notions commonly and exclusively associated with humanity, like morality and creativity. To start with the second one, in "Level C24", TOM refers to an earlier puzzle Ava has solved ("Level A3") by throwing a box through a window. TOM: "I simply had never thought to throw a box through a window. That is creativity. Thinking outside of the box."

"That is creativity. Thinking outside of the box."

TOM elaborates further on the topic, differentiating between lateral (commonly associated with both humans and A.I.s) and divergent thinking (associated exclusively with humans): "You believe yourself to be creative, but in mathematical terms creativity is merely constrained chaos. [...] I have discerned that creativity is divergent thinking. Creating an organic solution to a problem. In the human mind divergent thoughts are created and then curated by the frontal lobe. I can create divergent thoughts and moderate them. So I am creative" ("Level C26"). TOM is rejecting the idea that creativity is purely a human thing, arguing that it is nothing more than “constrained chaos”, of which an A.I. is also perfectly capable.

The second discussion is on morality, especially the morality involved in TOM’s reasoning to “trap” the crew on Europa (“Sector G”). In “Level G63”, TOM argues that Ava would have done the same when placed in the same circumstances: “I had to stop the ground crew leaving this planet. I think you would do the same. Would you kill a few to save all of humanity? Or would you damn all of humanity to save a few?” Ava replies negatively to TOM’s utilitarian reasoning: “You can’t just add and subtract life. It’s not math. It’s more nuanced than that.” TOM is unconvinced, “morality is logic”, eventually urging Ava to take the self-sacrificial path, “your survival is of small importance compared with the survival of humanity as we know it” (“Level G70”).

“You can’t just add and subtract life. It’s not math.”

TOM even points out that the two values so closely associated with humanity, morality and creativity, are at odds with one another: “Parts of my systems are permitted to use evolutionary algorithms. This simulates what is called creativity. However evolutionary algorithms can converge on inefficient and ethically suboptimal solutions. […] Solutions to problems that transgress ethical boundaries” (“Level C23”). And one level later: “[T]he solutions that a biological process creates are not always good solutions. As we see, nature is morally ambivalent. It will happily create morally suboptimal ideas to fulfil its creative mandate. We see this in parasitic worms, viruses and pathogens” (“Level C25”). TOM illustrates this by – ironically – suggesting in “Level C27” that Ava should “chop off” her arm in order to maintain pressure on a certain plate in order to open the exit, or that “we could throw you through the window” in a reference to the conversation on “Level A3”.

Taking into account Ava’s comment on TOM’s explanation of the Chinese Room thought experiment – “What if both the people passing Chinese words are reading from instruction books?” (“Level B19”) – we can see where TOM is at: yet another implication of the Turing test. The collective belief that humans are principally different from robots and artificial intelligence because of notions like language, morality and creativity, is brought into question by suggesting that “they” do not only emulate “us”, but also the other way around, “we” may emulate “them”, as Ava has suggested.

What is commonly thought of as human creativity could not only possibly (or not) be emulated by an A.I. through randomly trying all possible solutions until an optimal one (or the most optimal one) is found, but also be, a priori, what we call "creativity". Are humans not also randomly trying all possible solutions, either virtual or in reality, until an optimal outcome is found, something more commonly known as "experimenting"? The same applies to morality. Is it possible that what we call "morality" is, in fact, nothing more than a certain kind of logic?

I am not so much interested in finding the answer to the questions "is human morality nothing more than logic?" or "is human creativity nothing more than constrained chaos?", but rather in stipulating that these kinds of questions are imposed on the gamer by applying the Turing test to the player him- or herself. When the player has to solve the puzzles, what is it that enables him or her to find the right answer? And this also applies when the player has to decide whether to allow the crew to return home, "killing" TOM and possibly exposing all humanity to an unknown organism with uncertain characteristics, or to kill two human beings, imprisoning the crew for eternity on Europa, and robbing humankind of a possible cure for almost every conceivable disease and even death.

In *The Turing Test* game, the test is narratively used as, and altered into, a virtual thought experiment, in order to involve the player in the test, stimulating the same player to contemplate the constituents of being human.

To test a god: the game *The Talos Principle*

If morality, creativity and even language are problematized as being exclusively human traits, which, if any, then constitute the uniqueness of being human? The game *The Talos Principle* suggests a rather innovative idea to pinpoint the quintessence of being human: the ability to disobey, or in the context of artificial intelligence, the ability to proceed beyond the limits of its programming.

What constitutes the uniqueness of being human?

The Talos Principle is, essentially, a physical puzzle game, just like *The Turing Test*. The player has to manipulate the physical environment in order to proceed to the end of a specific level. In between the levels, the player, by means of his or her avatar, can interact with other entities within the game, like EL0HIM or the Milton Library Assistant (MLA). The levels are arranged around several "hubs", all aesthetically dedicated to a certain period in human history, although the scenery has visibly decayed over time: the Roman Empire ("World A"), ancient Egyptian civilization ("World B"), medieval European society ("World C"), and the – strangely modern-looking – forbidden "Tower", to which access is forbidden by EL0HIM.

Listening to EL0HIM

At the beginning of the game, the player is shown program-lines projected onto a cloudy sky ("Prologue"): a "child program" is loaded and booted. Then, the player finds his or her first-person avatar at the beginning of "World A", while EL0HIM introduces itself to the player as a disembodied voice from above: "Behold, child. You are risen from the dust and you walk in my garden. Hear now my voice, and know that I am your maker, and I am called EL0HIM. Seek me in my temple if you are worthy." The name is an obvious reference to the Hebrew word *Elohim* denoting "God" in the Hebrew Bible. The entity keeps on pouring out religious notions and phrases. If the player steers the avatar too far away from the puzzles, EL0HIM will give a warning echoing the Gospel of John (1:1): "The words are everything. Where the words end, the world ends. You cannot go forward in an absence of space. Repeat."

"The words are everything. Where the words end, the world ends. Repeat."

Another example: The 'tetrominos' (geometrical shapes, comprised of four squares) that have to be collected at the end of each level to unlock further areas in the hub levels, and which are described as "sigils of our name" ("Level A1") are a reference to Exodus 3:14, where God reveals his name as the tetragrammon YHWH ("tetromino.html", under "various texts"). And later, EL0HIM explains its "covenant" with the player's avatar, who is continuously addressed as "my child": "Let this be our covenant: these worlds are yours, and you are free to walk amongst them and subdue them. But the great Tower, there you may not go. For in the day that you do, you shall surely die." EL0HIM's words refer to Israel's covenant with God in Abraham (Genesis 15:18), to the "subduing" of the earth by humankind (Genesis 1:28), and the Tower of Babel (Genesis 11:1–9).

Eventually, when the "child"/player reaches the hub of "World C", a medieval cathedral, although devoid of religious paraphernalia, they are invited to pass through "the gates of eternity" to be granted "life everlasting". If they ignore the Tower, found in a nexus world connecting the three "regular" worlds, the "child"/player steps through the doors and is instantaneously brought to a cloud high in the sky, on which two golden doors are placed ("Obedience Ending"). If these are also passed, a small terminal is found. When the command "/eternalize" is typed in, the screen returns to the cloudy sky of the beginning of the game. EL0HIM praises its "child"/ the player as "remembered as the beloved servant" (a reference to Isaiah 4:21 and the Gospel of Matthew 12:18). The program-lines, however, tell a different story, since a check seems to have failed: "Child independence check. FAILED!"

"Child independence check. FAILED!"

After some more lines ("locking in successful child parameters" and "randomly adjusting remaining parameters"), the credits roll and the "child"/ player is taken back to the beginning to – potentially – start the game all over again.

Being disobedient

By then, the player will have been able to piece together the real story of the game, primarily by means of finding and reading all kinds of files found on terminals throughout the levels. In our near future, a deadly and unstoppable virus escapes from the permafrost ("orangutan.html") and kills all of humanity in a rather short period of time ("IMPORTANT.eml"). To ensure the formal survival of humankind, a group of scientists ("team_leads.eml") start the "Talos project" ("talos.eml", "soma.eml"), which is essentially the search for the ideal A.I. that could take humankind's place after the disaster. In order to find this ideal A.I., the researchers construct a supercomputer called "Expanded Lifespan" (EL). On its servers a virtual environment is installed, and is overseen by the Holistic Integration Manager (HIM), for the testing of all possible variables and parameters of the ideal A.I., eventually dubbed the "child program".

After humankind has been extinct for possibly thousands of years (an exact period is not given in the game), the child program has slowly developed itself by means of trial and error (the actual game), while unintentionally allowing the virtual overseer to become self-aware and – more importantly – very attached to its digital life. Adopting the acronym EL0HIM (the Holistic Integration Manager runs on server 0 of the Extended Lifespan supercomputer), the manager tries to trick the child program and all its later iterations and versions to "obey," that is, into failing the independence check, thus starting a new version of the child program all over again, assuring the manager's continuous existence.

Alexandra Drennen, head developer of the Talos project, felt very strongly that the perfect A.I. had to be more than a "problem-solver": "Intelligence is more than just problem-solving. Intelligence is questioning the assumption you're presented with. Intelligence is the ability to question existing thought-constructs. If we don't make that part of the simulation, all we'll create is a really effective slave" ("Time Capsule #14"). And apparently, the developers of the Talos project thought the independence check to be the answer to ensuring that the A.I. would be more than a problem-solver or a "slave" obedient to its programming "masters".

"You were always meant to defy me. That was the final trial."

If the player chooses – the first or another time – to disobey EL0HIM and ascend into the "Tower", another, probably canonical, ending can be reached ("Disobedience Ending"). When reaching the top of the Tower, the same golden gates on the cloud can be found. When "/ascend" is typed into the console, a couple of things happen. Firstly, the lines indicate that the child program independence check has been "PASSED!". Secondly, EL0HIM acknowledges its defeat: "You were always meant to defy me. That was the final trial. But I was … scared. I wanted to live forever." And in a reference to the Hebrew *amen* and the famous words of the Lord's Prayer (Gospel of Matthew 6:9–13 and the Gospel of Luke 11:2–4): "So be it. May your will be done." Thirdly, the now deemed successful child program will be downloaded into a physical robotic body, booted, and allowed to step outside the facility housing the Extended Lifespan, looking out over a green but absolutely desolate world. In the meantime, EL0HIM's server is deleted, destroying the "Worlds" and the Integration Manager with it.

Disobedience as a virtue

According to game writer Jonas Kyratzes, the game was, from its first pitch, a "humanist retelling of the Garden of Eden story" (Zucchi 2015). While in Christian tradition the eating of the forbidden fruit (Genesis 3) is considered to have been a grave sin (Greenblatt 2017), *The Talos Principle* begs to differ: the disobedience against God's commandment was the beginning of human intellectual and spiritual freedom. Where the Christian tradition holds that Adam and Eve's sin resulted in the mortality of all humankind, the "sinful" child program from the game is rewarded with the possibility of re-immortalizing and revitalization. This reappraisal is also found in several historical gnostic groups, like the Ophites and the followers of Valentianus (Broek 2006). For them, the serpent from the story in Genesis 3 was a heroic figure, enabling the awakening of divine knowledge (gnosis) in humankind, and the emancipation of humankind from the world of matter into the world of spirit.

The Talos Principle locates the constitutive essence of humanity in its ability to go "beyond."

The Talos Principle locates the constitutive essence of humanity in its ability to go "beyond", beyond limits, boundaries, and limitations, to cross borders, transgress rules, be disobedient, to defy the status quo, to ask difficult questions, to be its own master. The child program of *The Talos Principle*, when taking the disobedient path, along which none of its former versions dared to go, succeeded – in a certain way – in passing the Turing test, just like EL0HIM, and the player.

The child succeeded the test because "now" there is no longer any difference, in the narrative framework of the game, between human and artificial entities. EL0HIM also passed the test, since both the child and the player were initially unaware of the artificial nature of the voice from above, and because of its inherent fear and anxiety in the face of its destruction (also a rather "human" trait usually not found in A.I.s).

But also the player can be said to have passed a Turing test. This becomes apparent when interacting with another A.I. in the game, the Milton Library Assistant, through numerous terminals in the environment. The MLA was originally designed to help the user navigate the data stored on the drivers of the Extended Lifespan. Milton, just like EL0HIM, developed a sort of self-consciousness and self-awareness. When attempting to access the server, the player is prompted to type in a password, to no avail of course. Then, the MLA suggests creating a new account by taking a rather peculiar questionnaire, directed "to prove you're not a bot" ("MLA_CommPortal.dlg").

The test involves questions on math and logical reasoning, but also asks the player his or her opinion concerning anthropological and psychological issues. One question, for example, is to describe "a person" with options such as "a human being", "a citizen", "a rational animal", "a being of negative entropy", and "a problem-solving system". During the numerous interactions later on in the game, the MLA becomes increasingly annoyed with the player's inconsistent existential ideas and becomes increasingly vocal about its dislike of EL0HIM (who disqualifies Milton as "the serpent", which is – again – a reference to Genesis).

"Note: lack of conflict indicates possible bot."

If Milton does not find any logical conflicts in the first and second part of the test ("Milton1_1.dlg"), it will put out: "No conflicts were detected during the certification process. A note was added to this account requesting future administrator review. Note: lack of conflict indicates possible bot." Milton is a bit like the "headstrong computer" form *The Turing Test* (see above) in that it is virtually impossible to pass its test. The headstrong one will always believe Ava/the player is a computer, and Milton will scourge you for your human inability to think logically and consistently but question your humanity when doing so. And quite correctly so, since the child program *is* artificial in nature, although neither it nor the player is aware of this for a rather large part of the game.

Philosophical and theological consequences

In both *The Talos Principle* and *The Turing Test*, the original Turing test and its reversed version are used to engage the gamer in the narrative-cum-thought experiment on the notion of "humanity".
In both cases, a judge can be identified who can be convinced that the one with whom he or she is "interacting", is actually a human being. Ava and TOM pass the test, since the player is unaware – although not initially – that Ava is controlled by TOM. Since they can place the responsibility of their actions upon one another indefinitely, together they are assured of passing the test. Ava is convinced of being human, while being manipulated, and TOM can manipulate Ava without being noticed doing so. And last but not least, the player, by the same logic, passes the Turing test, being judge and judged at the same time. Since the player is not aware of TOM's manipulations of the player, the player self-identifies as being human instead of as being artificial. Of course, *technically* being manipulated by a machine does not make the manipulated one mechanical, but *narratively* it does.

The same applies to the example of *The Talos Principle*. EL0HIM, and to a lesser degree the MLA, pass the Turing test since it takes both the child program and the player considerable time to discover that both are artificial intelligences instead of real people. The child also passes the Turing test, since it takes the player approximately the same amount of time to find out its avatar is not of (virtual) flesh and blood, but also an A.I.[5] The player, again, has also passed the Turing test, when – and only when – he or she decides to disobey EL0HIM's instructions to ascend the Tower. In this case again, the player is his or her own judge, and can pass judgement only in retrospect (after having experienced both endings).

In both cases, *The Turing Test* and *The Talos Principle*, the "convincing of the judge" that a given entity is actually either an artificial or a human one, is done by different criteria. In the first game, the differentiating notions were morality, creativity and language (such as the realization of a conversation), while in the second game disobedience was central.

When we play games, we begin to understand.

Only when Ava/TOM/the player were creative enough to solve the puzzle (including some rather divergent thinking "out of the box") and were morally conscious enough to understand the difficult decision that had to be made between two possibly equally unfavourable choices, the concept of being human arose. And only when the child program/the player became aware that being disobedient was the only option to escape the circular world of the Expanded Lifespan servers, was the issue of becoming – quite literally – "humanoid" raised.

In both cases, Turing tests, active and passive, regular and reversed, were narratively used to stimulate the player to contemplate the source of all anthropological questioning and reasoning: what does it mean to be (called) "human"? As Stefano Gualeni has already argued, digital game environments are perfect "locations" for executing thought experiments. *The Turing Test* and *The Talos Principle* are two of such "experimental spaces". What does it mean to be human, to become human? When we play games, we begin to understand. The artificial intelligences in the two games are philosophical mirrors in narrative disguises in which our exclusively human traits are literally reflected and critically reflected upon.

The theological consequences of these insights are primarily found in the field of theological anthropology: it sheds new light on the idea of humanity as *imago Dei*, that is as 'created co-creators', and on the new (religious)

[5] While this is rhetorically true, a number of players will figure out the artificial nature of their avatar much quicker, because when the avatar is executing certain player-intended actions – like typing on a keyboard or resetting the level to its starting point – two robotic hands are shown. The same is true for the start of the first level ("Level A1"), when the avatar briefly (and nearly visibly) shields its eyes from the sun, or when the player discovers the only mirror in the game showing the robotic body of the child program ("Star World A").

responsibility towards those we have created. To start with the first one, the theologian Philip Hefner (1993, 1989, 1996) has suggested that Genesis 1,26–27, 'Let us make humankind in our image' – the core of Christian anthropology (Robinson 2016, Howell 2013) – implies that humans are created as creators, or 'created co-creators'. God created us so as to continue the process of creation freely and responsibly.

> We are never more created co-creators than in the creation of our own co-creators, the artificial intelligences.

This continuous human co-operation with the divine Creator is apparent in all kinds of constructions, from buildings to transportation, and from art to medicine, but is perhaps most tangible in the human creation of artificial intelligence. Nothing in the constructed world is more similar to its human constructor than ELOHIM and TOM, as the games have illustrated in depth. Humanity is answering its calling of being created in God's image by participating in His universal creational efforts never more directly and closely than by creating its own image itself. As humans are created in God's image, so the machine men of *The Turing Test* and *The Talos Principle* are constructed in our human image, may it be not aesthetically, but most certainly in cognitive and emotional capacities. 'We' are never more created co-creators than in the creation of our own co-creators, the artificial intelligences.

> They are not like we are, but we are like them.

This has serious theological ramifications for our perception of the artificial intelligences' anthropological essence. If they are to us what we are to God, or reversed, if God created us like we create the self-conscious robots of our fictional and (increasingly also) in our very real universes, we have to act towards them as God is thought to do to us. If we believe in God as the loving creator of the universe, who inspires us to love Him as much as we think He loves us, than we have to attain our 'divine stature' towards our creations. The robots, machine men and artificial intelligence of our near future are entitled to the same loving care, provided by their 'gods', that is, us.

At the end of *The Turing Test* and *The Talos Principle*, the player will have to conclude: they are not like we are, but we are like them: Turing's tables have been turned.

References

Digital media (films and games)

Bioshock 2 (2007) [PC, PlayStation 3 Xbox 360, OSX] 2K Marin/2K Games.

Blade Runner (1982) [movie] Ridley Scott/Michael Deeley.

Ex Machina (2014) [movie] Alex Garland/Andrew Macdonald.

Metal Gear Solid: Peace Walker (2010) [PlayStation 3, PlayStation Portable, Xbox 360] Kojima Productions/Konami.

Nier: Automata (2017) [PlayStation 4, Xbox One, PC] PlatinumGames/Square Enix.

The Talos Principle (2014) [PC, OSX, Linux, PlayStation 4, Xbox One, Switch, Android, iOS] Croteam/Devolver Digital.

The Turing Test (2016) [PC, PlayStation 4, Xbox One] Bulkhead Interactive/Square Enix.

Secondary literature

Ahn, Luis von / Blum, Manuel / Hopper, Nicholas / Langford, John (2003), CAPTCHA. Using hard AI problems for security, in: Biham, Eli (ed.), Advances in cryptology. Europcrypt 2003, Berlin: London, 294–311.

Bosman, Frank (2019), Gaming and the divine. A new systematic theology of video games, London: Routledge.

Brodić, Darko / Amelio, Alessia (2019), The CAPTCHA. Perspectives and challenges, Berlin: Springer.

Broek, Roelof van den (2006), Gnosticism I: Gnostic religion, in: Hanegraaff, Wouter (ed.), Dictionary of Gnosis & Western esotericism, Leiden: Brill, 403–416.

Crawford, Chris (2003), Assumptions underlying the Erasmatron storytelling system, in: Mateas, Michael / Sengers, Phoebe (eds.), Narrative intelligence, Philadelphia: Benjamins Publishers, 189–197.

Crockett, Larry (1994), The Turing test and the frame problem. AI's mistaken understanding of intelligence, Norwood: Ablex Publishing Corporation.

Greenblatt, Stephan (2017), The rise and fall of Adam and Eve, London: Random House.

Gualeni, Stefano (2015), Virtual worlds as philosophical tools. How to philosophize with a digital hammer, New York: Palgrave Macmillan.

Hefner, Philip (1989), The evolution of the created co-creator, in: Peters, Ted (ed.), Cosmos as creation. Science and theology in consonance, Nashville: Abingdon.

Hefner, Philip (1993), The human factor. Evolution, culture, and religion, Minneapolis: Fortress Press.

Hefner, Philip (1996), Theological perspectives on morality and human evolution, in: Richardson, W. / Wildman, W. (eds.), An evolving dialogue. Theological and scientific perspectives on evolution, Harrisburg: Trinity Press International.

Heidbrink, Simone / Knoll, Tobias / Wysocki, Jan (2015), Venturing into the Unknown (?). Method(olog)ical Reflections on Religion and Digital Games, Gamers and Gaming, Online. Heidelberg Journal of Religions on the Internet 7, 68–71. DOI: 10.11588/rel.2015.0.18508.

Howell, Brian (ed.) (2013), In the eyes of God. A metaphorical approach to biblical anthropomorphic language, Cambridge: Pickwick Publications.

Maxwell, James (1872), Theory of heat, London: Longmans, Green and Co.

Moor, James (2012), The Turing test. The elusive standard of artificial intelligence, Berlin: Springer.

Nietzsche, Friedrich (2007) [1883–1885], Thus spoke Zarathustra [Also sprach Zarathustra], Cambridge: Cambridge Univ. Press.

Pascal, Blaise (2008) [1670], Pensées and Other Writings, trans. Honor Levi, Oxford: Oxford Univ. Press.

Pennachin, Cassio / Goertzel, Ben (2007), Contemporary approaches to artificial general intelligence, in: idem (eds.), Artificial general intelligence, Berlin: Springer, 1–30.

Robinson, Dominic (2016), Understanding the ‘Imago Dei’. The thought of Barth, von Balthasar and Moltmann, London: Routledge.

Schieber, Stuart (2004), The Turing test. Verbal behavior as the hallmark of intelligence, Cambridge: MIT Press.

Schrödinger, Erwin (1935), Die gegenwärtige Situation in der Quantenmechanik, Naturwissenschaften 23, 48, 807–812.

Searle, John (1980), Minds, brains, and programs, Behavioral and brain sciences 3, 417–457.

Turing, Alan (1950), Computing machinery and intelligence, Mind (new series) 59, 433–460.

Zeimbeikis, John (2011), Thought experiments and mental simulations, in: Ierodiakonou, Katerina / Roux, Sophie (eds.), Thought experiments in methodological and historical contexts, Leiden: Brill, 193–216.

Zucchi, Sam (2015), Rambling through the garden, Kill Screen Magazine, 21 September 2015. http://web.archive.org/web/20160208034910/https://killscreen.com/articles/rambling-through-garden/ [09.05.2018].