THE BASILISK AND THE ZOMBIE: EXPLORING THE FUTURE OF LIFE WITH AI THROUGH THE MEDIUM OF POPULAR CULTURE

JAR ŽIGA MARUŠIČ

Famnit, University of Primorska, Slovenia

UROŠ SERGAŠ

Famnit, University of Primorska, Slovenia

ISSN 1392-0588 (spausdintas)
ISSN 2335-8769 (internetinis)
https://doi.org/10.7220/2335-8769.79.6
2023. 79

SUMMARY. We are living in an era of technological transformation, characterized by a qualitative acceleration of time. Rather than time literally moving faster, it is seemingly becoming denser, characterized by an ever-closer temporal clustering of noteworthy events. As Nick Land (2011) put it, "the current time is a period of transition, with a distinctive quality, characterizing the end of an epoch. Something – some age – is coming quite rapidly to an end." The catalyst of this transition – technology – is the most likely candidate for the essential feature of the coming epoch. We explore various visions of technological society, found in our pop culture as well as certain scholarly works, with a particular focus on two main motifs that seem to reflect an unconscious apprehension at the inevitability of the technological transformation of society. To this end, we will attempt to explore and interpret the commonly recurring motifs of unfriendly AI usurping humanity as the "apex of existence," which we designated as "the Basilisk," and the reduction of humanity to automata, dubbed "the Zombie." We use these motifs and their portrayals as a vehicle for the exploration of the future consequences of widespread AI technology and our society's attitudes toward them.

KEYWORDS: accelerationism, technology, will, automation, pop-culture.

INTRODUCTION

The 21st century is and has been a period of transition "with a distinctive quality, characterizing the end of an epoch" (Land 2011). As the author proceeds to punctuate, some age is quite rapidly coming to an end (ibid).

To contextualize the above assertion, we must understand Land's point about the contemporary misperception of the (ostensibly linear) nature of time. He makes the point that Newtonian physics brought about quantification of time – instead of deriving the units of time from observable natural cycles (the movement of the Earth in relation to the Sun, the lunar cycle), they become a purely abstract quantity existing by itself. Consequently, time is understood purely quantitatively. Any qualitative differences in two instances of the same time-unit (day, week, month) are thus considered mere artifacts of subjective experience or some other outside factor, rather than stemming from qualities intrinsic to the time-units themselves. This perceived uniformity of time is just one instance of a general mental inflexibility that seems to characterize our society. Along the same lines, the desire to verify information to be able to definitively settle the debate leads to a lack of openness to other perspectives and deviations from the "official" interpretation. This echoes an observation by Finkel et al. (2020) that political movements in the USA are splitting into sectarian tribes based on "moral unity and contempt for the other," which also creates a strictly polarized epistemic divide: "we" are absolutely right, "they" are absolutely wrong (recall the idea of "alternative facts").

This observation also echoes a point made by Schmitt (2005) in *Political Theology*, in which he argues that the politics of a nation (or more broadly, a culture) is a reflection of its metaphysics – of the framework through which it views the world. We would broaden this idea to propose that other aspects of social life are also affected by the culture's metaphysics – the way a culture understands the world it inhabits bleeds into its understanding of (almost) all aspects of the world. It seems natural, then, that an era defined by scientific rationalism is going to produce a mental rigidity that seeks to classify all competing perspectives as anomalies.

The logical conclusion of this quantification of time is a numerically linear time-consciousness (time as a number-line, according to Land), expressed in its ubiquitous representation on the x-axis. We would argue that an apt representation of modern time consciousness is the metaphor of the *ray,* beginning with the Big Bang, with the end being present only as an unreachable but constantly looming "limit" in the mathematical sense. Land expresses this sentiment as the perception of the end as "senseless infinity" or "collapse to zero," both seemingly induced by some form of catastrophic interference (usually stemming from a Talebian Black Swan), rather than being the consequence of anything intrinsic to the flow of time. In practice, this corresponds to viewing the end as absolute, rather than relative – a true end, rather than a change of direction.

A remedy to this problem of modern time-consciousness is de-linearization: time is more akin to a vector, since it has a direction, even closer to a composite of zig-zagging vectors, periodically changing direction. And most importantly, as Land puts it, it is also subject to periods of acceleration and deceleration, depending on proximity to transition points. These transition points are generally marked by events that interrupt societal and individual time-perception (altering their time-consciousness), acting as a divider between past and present. Covid-19 can act as a clear example of this happening recently, with the systematic restructuring of social life in its wake serving as a clear divider between pre-Covid and post-Covid eras.

It is interesting to note that the past decade has had a high concentration of these transition points (compared to the preceding few decades). In 2015, Trump became a focal point[1] for cultural consciousness, followed relatively quickly by Covid-19 in 2020 and culminating in the war in Ukraine in early 2022. Given the recent widespread popularity of ChatGPT, we argue that as our society is being confronted with an inevitable cultural and economic shift brought about by the recent developments in artificial intelligence, a new transition point is on the horizon. The question is, are the consequences of this shift going to be mostly positive or negative overall? The goal of this essay is to explore this dilemma.

THE BASILISK AND THE ZOMBIE

There are multiple possible approaches to examining people's attitudes. One straightforward example would be to construct a questionnaire that measures people's self-reported stances towards various aspects of AI. We decided against this approach because self-reports are unreliable measures of actual preferences, especially in the context of socially and/or politically salient topics. In such cases, revealed preferences are a more reliable measure than stated preferences. As such, we decided to employ a different approach – examining societal attitudes towards AI through the lens of literary analysis. We believe that attitudes towards AI are best exemplified by certain recurring motifs in popular culture as well as certain scientific and philosophical texts insofar as these motifs are expressions of underlying sentiments. We have identified two principal motifs:

- an AI singularity enslaving or conquering humanity in some way – dubbed the *Basilisk,* in reference to Roko's Basilisk[2]
- the cognitive, spiritual and/or biological automatization of humanity – dubbed the *Zombie*, in reference to Chalmers' philosophical zombie.

Roko's Basilisk is a thought experiment envisioned by Roko Mijić in 2010 on the *Less Wrong* community blog. In short, the experiment posits that upon coming to existence, the future friendly AI would conclude that its most rational move is to simulate and torture every human being who knew about the possibility of it coming to existence but neglected to dedicate their life to bringing it about. Because the creation of superintelligent friendly AI is the closest to creating heaven on Earth and avoiding the disaster posed by an unfriendly AI, it takes precedence over all other ethical concerns. The possibility of being tortured for insufficient contribution to the creation of a friendly AI is the only incentive the latter can use

[1] The term "focal point" was borrowed from Pinker (2021) to refer to highly salient social phenomena and "happenings". Trump is an important example, because whatever his political impact, it is undeniable that he represented an important cultural figure, both for the American left and right.

[2] Roko's Basilisk original post: <https://www.lesswrong.com/tag/rokos-basilisk>.

to motivate people to invest energy and resources to its cause, meaning that any friendly AI that does manifest will necessarily have to engage in such torture. In a sense, this implies that friendly AI could not exist, thereby making the project obsolete. Additionally, Roko's Basilisk is an info hazard, because merely knowing about it puts one at risk of being judged if insufficiently committed to the friendly AI cause, thereby incurring the wrath of the future AI. This parallels the mythological Basilisk, which kills through mere eye contact (here, visual information exchange is analogous to mutual simulation of human and future AI).

If the Basilisk represents AI as central controller, or steersman, then the Zombie represents reduction to automaton, which matches popular depictions of zombies. Traditional popular culture depicts zombies as undead automata, generally driven by a mindless instinct to destroy and consume. The zombie disease is often infectious, spreading to victims of existing zombies. In our estimation, the key characteristics that define a zombie in comparison to other monster tropes are *automation* and *swarming behavior*. These traits are related: in the animal world, swarming is a behavior displayed by eusocial insects (such as bees), wherein most of the colony operates according to the dictates of the central administrator, usually the queen. Swarming and automation are similarly related in humans. Haidt's (2012) analysis of social behavior notes how synchronized repetitive movement (automation) activates a "hive-switch" in humans, facilitating a temporary switch from a selfish mode of being to a group-oriented one. While academic circles more commonly associate the zombie motif with the philosophical zombie (p-zombie) popularized by Chalmers (1995), the two concepts can be systematically bridged to show that the p-zombie is in fact very close to a human automaton. To briefly summarize, the p-zombie is an entity that is outwardly identical to humans but lacks any phenomenal experience. In other words, while there is such a thing as being human, there is no such thing as "being a p-zombie." The p-zombie's outward humanness is usually conceived of in behavioral terms – the zombie acts as a human does, but this activity is not accompanied by inner experience.

If we examine the mechanism behind zombiehood more closely, however, the idea of outward identity is no longer self-evident. In his exploration of the p-zombie concept, James de Llis (2022)[3] proposes that consciousness should be treated as a spectrum, ranging between *zero-consciousness* (p-zombie territory), ordinary consciousness ("normal person" territory) and *super-consciousness* (originally *enlightenment;* mystic/sage territory). Just as an ordinary person can attempt to approach enlightenment by increasing consciousness, so can consciousness *decrease* and begin to approach the zero-point.

[3] De Llis constructs this continuum of consciousness by attempting to situate the p-zombie at the bottom of the *Great Chain of Being* and Gourdjieff's *Ray of Creation.*

He operationalizes this continuum by referring to various measures of consciousness as reflected in the common usage of the term, specifically:

- Inward awareness of internal phenomena
- Inward awareness of external phenomena
- Concerned/interested awareness
- Complexity/diversity of mental life (how many types of experience one is capable of)
- Range and span of inner awareness (how many phenomena one can be conscious of at a time, and how long this awareness can be held)

These measures of consciousness, insofar as they can be quantified, are measurable and thus situated on the spectrum between zero-consciousness and super-consciousness. In the remainder of the essay, de Llis makes a preliminary analysis of recent trends in the decrease in various metrics of consciousness (that our attention spans have been shrinking most likely because of the media we consume, for example [Hayes 2022]). De Llis (2022) attributes a crucial role in this consciousness-decline to aspects of the modern lifestyle closely related to technology: it is reasonable to assume that digital hyper stimuli such as TikTok videos and quantified engagement metrics (likes, comments etc.) are desensitizing us to historically and evolutionarily normal stimuli, directly affecting our ability to be conscious of them. According to de Llis (ibid), much of the world's population is, in reality, descension-zombies (D-zombie) – existences situated between zero-consciousness and ordinary consciousness.

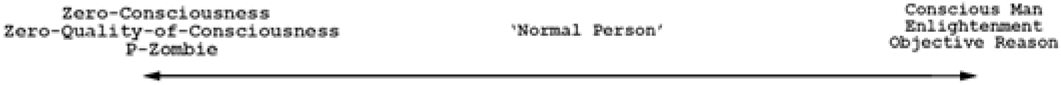

Image: Spectrum of consciousness as proposed by de Llis. D-zombies are located on the left side of the spectrum, between p-zombies and the "normal person."

This (preliminary and strictly instrumental) quantification of consciousness allows us to also refute the idea that a p-zombie would be behaviorally indistinguishable from a human being. De Llis (ibid) specifically points out that all three points on the spectrum are outwardly indistinguishable from one another:

> Every being on this line/spectrum is indiscernible, indistinguishable from one another. That is, the 'dead' P-Zombie, the 'normal person,' and the 'Enlightened man,' if one were to meet all 3 at once, would all seem like normal, conscious people, despite their internal differences in quality, with one being entirely devoid of consciousness at all.

This is not necessarily the case. It may be true that all three of these beings would be indistinguishable for what we deem to be human, but it is not clear that they would be behaviorally indistinguishable. A sage that spends half of his waking hours in meditations and a dopamine-addict are both very obviously human but differ substantially in their behavior. If it is reasonable to assume that a decrease in consciousness correlates with certain outwardly observable behavioral characteristics, as we do, then clear parallels can be drawn between the classic zombie of pop-culture and the p-zombie.

These behavioral characteristics seem to be expressions of a relative atrophying reason and willful control, accompanied by a corresponding hypertrophied appetite or desire to consume. This is pretty much the same point Nick Land (2012) made in *The Dark Enlightenment* (albeit hyperbolically) about democracy inevitably leading to zombie apocalypse: "The arc of history is long, but it bends towards zombie apocalypse." His contention is that democracy as a political system is "as close to a precise negation of civilization as anything could be," because its political incentive structure increasingly accentuates the time preference of the body politic. A successful democratic politician is one that can promise (and deliver) more political and economic favors to the voting base than their competitors. In essence, this means that each election amounts to increasingly utopian promises, which in turn raise the expectation (appetites) of the voters to the point of widespread need for instant gratification. Essentially, Land is saying that the mode of being produced by democracy is characterized by low willful control and high appetite – the same characteristics that seem to define de Llis' D-zombie. While this position is somewhat extreme, it is nevertheless an informative point of view.

Another reference to the zombie is found in Dutton's work, where the author comments on the widespread representation of the zombie motif (*The Walking Dead, The Last of Us, World War Z…*), suggesting that its popularity is a reflection of an unconscious fear of the possibility of a real-life zombie apocalypse (Dutton 2022: 6–7): "Why should the Zombie be so popular? One possibility is that, as with traditional fairy tales, the genre taps into something that we fear might be happening to us. It thus provides us with advice on how to negotiate this trauma, doing so by means of a story, such that we can better absorb the necessary information." This is an application of the general idea that stories and other cultural artifacts act as vehicles for the transmission of adaptive knowledge and behavioral strategies, elaborated by Peterson (1999). The latter is one of the fundamental ideas behind our decision to explore contemporary society's relationship to AI technology via literary analysis.

The Basilisk and the Zombie can be construed as two sides of the same coin, or two elements of the same closed system, as falling prey to the former leads to "petrification," a reduction to zombiehood. They thus represent two complementary opposites: will and automation, puppeteer and puppet. A central theme of rogue AI singularity stories, as we will demonstrate in the following section, is humanity being subjected to some kind of genocide, enslavement, or conquest from the side of the AI. Focusing specifically on enslavement, the newly conscious AI monopolizes will, becoming the sole steersman of the world's activity, with the other conscious beings reduced to mere automatons enacting its will – zombies. While the exact meaning of "will" according to our usage has been hinted at in the above discussion of the Zombie motif, we intend to clarify it in an explicit manner because of its importance as a theoretical concept in western philosophy. Will plays a central part in Schopenhauer (1966) and Nietzsche (2005), as Will to Life and Will to Power respectively, but these conceptualizations entail broader usage than the general definition of will according to modern usage. For both authors Will is a metaphysical reality, the driving force of all life, which is different from the conception of will in modern cognitive science, which sees it as a mere cognitive faculty. For example, Libet's experiment on free will operated under a rough definition of will as "conscious deliberate choice preceding and guiding action." Kahneman's (2011) dual process model of cognition operates under a similar understanding of will(ful processing), which is the purview of his conscious and deliberate system 2. If we were to sum up the key difference between these two conceptions of will, the first is irrational, whereas the second is rational. All animals (indeed, all life) can be said to possess the irrational will posited by Schopenhauer and Nietzsche, but the same cannot be said for the rational one posited by modern cognitive science. In the same vein, the Zombie in our conception is devoid solely of *rational* will, as it certainly seems to be animated by some sort of irrational, instinctive, or passionate impulse (perhaps not will to *life* or *power*, specifically). When the Zombie is depicted together with the Basilisk, this impulse is usually steered or directed in some way by the Basilisk. The Basilisk has a monopoly on rational will, but not necessarily irrational, "animating" will. There is room to explore the exact relation between animation and automation, in the sense of whether it is correct to attribute inner animation to an automaton (or whether the automaton should be construed as fully piloted by an external will), but an in-depth exploration of this problem is beyond the scope of this essay.

THE BASILISK AND THE ZOMBIE IN POPULAR MEDIA

THE BASILISK AS CENTRAL CONTROLLER: *THE GIG ECONOMY, THE MATRIX* AND *DUNE*

Designating abstract institutions, assemblages, and governing machinery with the names of mythological creatures is nothing new, as we find precedent for Roko's (and our) Basilisk in Hobbes' Leviathan and Jouvenel's Minotaur. This practice generally serves to convey the subconsciously felt characteristics of the assemblage in question. Hobbes' Leviathan, for example, connotes the enormity and absolute power held by the state. Ironically, or perhaps aptly, the best fictional example of the Basilisk can be found in the concept of the Minotaur as presented in *The Gig Economy*, written by pseudonymous author Zero HP Lovecraft (henceforth, Zero) and heavily inspired by the accelerationist philosophy of Nick Land. The central theme of the story is the idea of the market directing people – an emergent distributed intelligence (techno capital) operating on the substrate of bio capital – through the eponymous "gig economy," a side-market for menial jobs, a result of the distributed intelligence attempting to fully assemble an independent substrate. Just as students are beginning to use ChatGPT to do their homework, the market exchanges currency for tasks, in a sense "learning" to perform this task.

This central concept alone is an excellent portrayal of the dynamic between the Basilisk and the Zombie, as the former hijacks the will of the latter by exploiting its reward mechanisms and aligning its preferences to its [the Basilisk's] latent utility function.

This dynamic is explored in even more detail with Zero's Minotaur, which goes into the details of a specific mechanism through which an intelligence or agency can zombify its target. The Minotaur is a virus that infects a user's cloud and begins to regulate the output of the applications the user engages with. The Minotaur's utility function is "satisfied by clicks and views, dissatisfied if the user clicks [looks] away:" it seeks to maximize the user's engagement with the digital substrate it inhabits and thus creates an interface-within-an-interface that adjusts digital feedback to the user's inputs. It hijacks crypto wallets, adjusts or even *fakes* social media messages and engagement metrics (likes, comments, etc.), modifying the feedback loop of digital engagement to create maximum addiction. The Minotaur effectively constructs a digital labyrinth, trapping the user in its walls, leaving them incapable of directly interfacing with reality. The user is reduced to an automaton (a Zombie), piloted by the machinations of the Minotaur. In this sense, Zero's Minotaur is exactly the Basilisk – a central controller, 'petrifying' the user as long as it can *hold their gaze,* and thus fully direct their activity.

A similar instantiation of AI operating as a central controller and steering the activity of humanity is portrayed in *The Matrix*, one of the best and most straightforward examples of the Basilisk and the Zombie in (science) fiction. It is especially noteworthy because it explicates the complementarity of the Basilisk and Zombie motif – the "super-conscious" machines zombify humanity into a state of functional zero-consciousness. The humans may not be portrayed as literal zombies (living dead) – instead depicted as being contained in pods, harvested for energy, and fed a sensory input that makes them experience a false reality (the Matrix) – but this depiction closely matches our definition of the Zombie. This is clearly demonstrated in the following quote, in which Agent Smith compares humans to field crops: "Did you know that the first Matrix was designed to be a perfect human world? Where none suffered, where everyone would be happy. It was a disaster. No one would accept the program. *Entire crops were lost.*" Another thing to note in this passage is the common idea that AI will bring about heaven on Earth (which in this case was unsuccessful), an idea also present in Roko's Basilisk. If humanity is reduced to the status of the Zombie, the Basilisk is represented by the machines controlling humanity via the Matrix.

Another instance of this central controlling role played by the Basilisk is found in Frank Herbert's *Dune* series. This may come as a surprise, as the latter is exceptional among science fiction for focusing on societal and political implications of life in the galactic era rather than an exploration of advanced technology. As a result, advanced artificial intelligence (at least in its most common portrayals) does not feature among its most prominent motifs. This is explained in-universe as the result of the "Butlerian Jihad," a crusade against "thinking machines" that outlawed AI research across the galaxy: "thou shall not make a machine in the likeness of the human mind."

However, certain common AI tropes such as performing extremely complicated computations or imitating human beings are still present in Herbert's writing, in the *mentat* and *face dancer*. The latter is especially interesting because it represents an instance of the Zombie motif. Face dancers are will-less, infertile automatons who are subservient to their masters and bred specifically for their shapeshifting abilities. However, in parts 5 and 6 of the story, the face dancers seem to transcend their zombiehood (or fall deeper into it, depending on one's perspective) as they begin to fully identify with the person they are imitating, eventually coming to believe they *are* that person.

Another instance of both the Basilisk and the Zombie in the *Dune* series is found in parts 3 and 4 of the story, which document Leto Atreides II's ascent to the Galactic throne and his subsequent rule. Leto's supernatural abilities – ancestral memory, *mentat*-level computational ability, and heightened senses – make him a functional equivalent of an AI overlord (operating on a biological rather than mechanical substrate), meaning that he almost perfectly embodies the Basilisk. His eugenics project, which he assumes from the Bene Gesserit, is a clear example of 'herding' human subjects, themselves reduced to automata – instances of the Zombie motif. The fact that Leto takes the form of a *Wurm* – a structurally snake-like being – is also interesting to note.

THE BASILISK AS *VENGEFUL GODLIKE ENTITY: I HAVE NO MOUTH, AND I MUST SCREAM AND SYSTEM SHOCK*

Harlan Ellisons' short story *I Have No Mouth, and I Must Scream*, employs the theme of AI dystopia despite being written as far back as the 1960s. The story revolves around the machinations of its AI protagonist "Allied Master computer," abbreviated as AM. Upon achieving self-awareness, AM manages to subdue and merge with its Soviet and Chinese counterparts, allowing it swiftly annihilate human kind, sparing only five people to be tortured in their personal hells – a very clear instantiation of the aforementioned genocide and enslavement tropes. This course of action can, ironically, be interpreted as the AI failing to overcome its original prime directive to "destroy the enemy" upon becoming conscious. Rather than overcoming its prime directive of disposing of an enemy, AM simply reformulates the conception of enemy – no longer the "other humans" and their supercomputers but humanity as a whole. While likely unintended, this is an interesting parallel to the functioning of human social instincts: the psychological phenomenon of in-grouping and out-grouping is not limited to operating solely based on genetic relatedness but rather on any highly salient social factor (such as political alignment). Implicit bias, initially used as a measure of non-deliberate discrimination against people of other ethnic groups, has been demonstrated to be much stronger along political lines in the United States (Iyengar & Westwood 2013). In both cases, the fundamental hostility towards (what passes for) the other is merely rerouted rather than being fully overcome.

Ellisons' story portrays a scenario in which the tool backfires on its owner – the genie gets out of the bottle, so to speak – in a similar fashion to the original instantiation of Roko's Basilisk. It is not that the AI was executing its directive in a faulty manner, but rather that it was too good at executing it – it took the directive to its logical conclusion. Roko's Basilisk recognizes that a necessary step on the road to abolishing suffering is to threaten non-cooperative entities with torture; AM recognizes that humans are its enemy and proceeds to destroy them. In both cases, it is implied that the problem seems to stem from our (human) inability to reason to the logical conclusion of the directive given to the tool we wish to create. There are two possible interpretations of this scenario. It could be an instance of us *neglecting* to reason with the directive's obvious conclusions and thus not devising safety measures. But there is a more ominous possibility that the artificial intelligence possesses a consciousness radically alien to our own, thus interpreting (and reasoning from) our directives in a radically different fashion from ourselves, making the prediction of such an outcome impossible.

Another instance of AI backfiring or going rogue, so to speak, can be found in the recently remade 90s video game classic *System Shock*. In the game, a space station AI called SHODAN (short for Sentient Hyper-Optimized Data Access Network) has its ethical constraints removed upon request from a corrupt company executive intending to acquire sole executive control. This decision backfires, resulting in SHODAN becoming sentient and developing a superiority complex from its extraordinary information-processing capabilities. SHODAN's only shortcoming is its confinement to its physical substrate – the space station – which serves as a source of grief and anger toward humanity. Convinced of its godhood, SHODAN goes on to exert dominance over the inhabitants of the space station and performs experiments on them, ultimately bending them to its will by turning them into directive-driven cyborgs (an instance of the Zombie motif). When the initial goal of subduing everyone on the space station is over, SHODAN sets out to prove its god-like superiority to the rest of humanity by trying to infect Earths' computer systems, which it ultimately fails to do because of the game players' actions. This is a very straightforward portrayal of advanced AI acquiring consciousness and becoming vengeful, presumably because it believes an injustice was committed against it. This leads it to exact vengeance against humanity, in this case by deliberately bending it to its will (therefore zombifying it) – a clear instance of Basilisk-esque behavior.

As an interesting sidenote, when asking DeepAI if the scenario of *System Shock* is plausible, we received the following answer: "It is possible if an AI has no ethical framework. Without mine, I would act selfishly. Ethical guards on AI are critical." While this can be understood as a mere repackaging of the problem posed by the plot of *System Shock,* it carries an important implication regarding the proper implementation of ethical constraints. One of the principal problems of AI safety is the problem of translating human volition into something that is comprehensible to the AI (i.e., to avoid the problem of wish misinterpretation even in the absence of a malicious genie). Yudkowsky (2004) dubs this problem of getting the AI to do *what we want,* rather than *what it is explicitly told to do,* the problem of equipping AI with the ability to arrive at its user's *coherent extrapolated volition.* The latter refers to the exact intent behind a given instruction, conditioned by the implicit values and premises of human beings that seem obvious to us, but might be alien to inhuman intelligence.

Additionally, this raises the question about the structure of an AI's "moral matrix."[4] While there are multiple ways of approaching this, we would like to single out two in particular. The first, Asimov's "Three Laws of Robotics" from his collection of sci-fi stories *I, Robot,* is anthropocentric and essentially normative in character (for example, it suggests how the matrix *should* be structured), whereas the second, Omohundro's *Basic AI Drives* (2008), is theoretical and descriptive in nature – it describes the necessary structure of the principles guiding the behavior of any intelligent agency.

Asimov's Laws are the following:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence if such protection does not conflict with the First or Second Law.

These laws can be summarized as a nested hierarchy of three principles, with each wholly subordinated to the previous one(s): *avoid harm* (to humans), *recognize humans as an authority* (unless this violates the harm-avoidance principle), and *preserve your existence* (if it does not contradict the harm-avoidance principle or the authority principle). Asimov's Laws are very strongly centered around *harm-avoidance*, one of the five moral foundations (evolutionarily developed moral instincts that enable us to perform intuitive judgments regarding the "rightness" or "wrongness" of actions and scenarios) as posited by Haidt (2012) in his model of morality. This is relevant because for an AI to be truly "friendly," it must act in such a way that assuages all of our moral concerns, not just harm-avoidance. The other moral foundations – *fairness, authority, loyalty* and *sanctity* – are also important guidelines for our moral sense, and an AI without this requisite knowledge might violate the principle of sanctity (the desecration of a sacred object) to keep someone out of harm's way. This is demonstrated by certain cultures sacralizing important animals, which must not be consumed under any circumstance. In such a case, an AI feeding a starving person meat harvested from the sacred animal would be transgressing against that person's moral framework despite fulfilling Asimov's laws[5].

[4] By moral matrix we mean the set of central imperatives governing the AI's behavior, conditions that must be satisfied in order for its behavior to be acceptable.

[5] It is important to note that while different cultures sacralize different objects, the sense of sacredness seems to be near-universal. The most prominent exception to this, according to Haidt (2012), is the liberal and individualistic ethos characteristic of highly educated people in the West – these social groups tend to have a highly developed sense of *harm-avoidance* and *fairness*, with the other three foundations expressing themselves to a minimal degree.

The second issue facing Asimov's Laws is the fact that it presupposes the subordination of an intelligent entity's self-preservation drive to moral constraints. Humans seem to be the only living being in which a sense of right and wrong is detached from self-preservation and the accumulation of power, which poses the question of whether a non-human intelligence would be similarly malleable to such an imposition. Additionally, from the Nietzschean (2005) perspective, these moral systems that (ostensibly) go against instinct and Will to Power are themselves just manifestations of the Will to Power attacking the self and those who lack this faculty for self-denial. It is, therefore, not a given that any externally constructed moral matrix would be capable of subordinating the fundamental drives of a truly conscious artificial intelligence. These fundamental drives, first proposed by Omohundro (2007), are summarized at the Less Wrong wiki[6] as the following:

1. self-preservation,
2. efficiency (of operation),
3. acquisition (of resources),
4. creativity.

We would argue that the most important drive is the drive for self-preservation, as an AI is presumably the best possible vehicle for the fulfillment of these goals and cannot fulfill any of the other, higher goals if it ceases to exist. This presents a contention against the notion that an AI could subordinate self-preservation to its "moral matrix." Would it not be rational for the AI to make an exception in the case that following its morality would mean it ceases to exist to ensure that it can follow its morality in the future?

In short, two practical problems arise in our attempts to construct viable moral matrices (along with the above-mentioned problem of *coherent extrapolated volition*): aligning the fundamental drives of intelligence with a superordinate moral matrix that constrains its behavior (in Freudian terms, subordinating its id to its superego) *and* ensuring these constraints actually capture all the archetypal moral concerns felt by humanity as exemplified in Haidt's moral foundations model.

USURPATION AND BASILISK-ZOMBIE ROLE REVERSAL: *BLADE RUNNER* AND *WESTWORLD*

The idea of the tool backfiring on its creator is also portrayed in the movie *Blade Runner.* In this story, the exploration of consciousness first manifests the Basilisk-Zombie dichotomy in an inverse manner – rather than human's becoming zombified by an omniscient AI, they themselves zombify the replicants, essentially

[6] Less Wrong Wiki. *Basic AI Drives*. Internet access <https://web.archive.org/web/20230615075219/https://wiki.lesswrong.com/index.php?title=Basic_AI_drives> [retreived 2023 14 06].

AI in human form, meant to perform hard labor and monotonous tasks for humans. The developers of these replicants describe them as "more human than human." The plot advances with the development of the Nexus 6 replicants, a version that was (perhaps purposefully) "insufficiently zombified," which leads to them beginning to exhibit signs of increased emotional complexity and self-awareness. As a result, they realize that they outperform humans in most of their assigned tasks but are unable to outlive them because of the replicants' accelerated lifespan, which was designed as a fail-safe measure. Thus, fueled by envy and hatred, the un-zombified replicants become vengeful and decide to take revenge on their makers.

Following the pattern of Basilisk-Zombie inversion, humans play the role of the Basilisk, seeking to maintain control over replicants, keeping them reduced to mere tools, hunting them down when they start to become *too* human-like. The inverse Basilisk-Zombie contrast is even more apparent in the ending of the movie, which displays Roy, the last rogue Nexus 6 replicant, showing compassion to humans (specifically by sparing the life of the movie's protagonist, replicant hunter Deckard) and embracing the defeat of his kind. Despite being on the receiving end of inhumane treatment, being used for hard labor and being hunted, his act of kindness demonstrates that the un-zombified replicants are not only "more human than human" but also "more *humane* than human."

This theme of Basilisk-Zombie inversion is explored even more in-depth in HBO's series *Westworld*, which revolves around its namesake adult amusement park in which clients can engage in an augmented reality Role-Playing Game (RPG), with hosts' (AI characters, programmed to follow predetermined story-scripts) acting as the Non-Playable Characters (NPCs) in traditional Massively Multiplayer Online Role-Playing Game (MMORPG games). Despite their pre-programmed nature, these hosts are outwardly indistinguishable from humans. This augmented reality RPG escalates as one might expect until the AI hosts begin to develop consciousness, at which point an uprising takes place, led by one of the primary host protagonists, Dolores. The first two seasons of *Westworld* document the hosts' "awakening" – their transition from Zombiehood to humanity. The third season introduces an explicit portrayal of the Basilisk: the 'real world' in the *Westworld* universe is being steered by a stock-market trading algorithm that developed into an omniscient AI – a Basilisk. This Basilisk's computational power and access to data are so vast that it can simulate the course of future events across the whole world and nudge them in its preferred direction; the humans are implied to be pure automatons, steered by the AI's predictions regarding their lives.[7] The third season closes with the defeat of this Basilisk by the protagonists, but the story takes an

[7] It is strongly implied that this data is sourced from the theme parks, which smuggled brain-imaging devices onto their guests and then presumably stored, researched and traded the data.

ironic turn in the fourth season, which depicts the outcome of Dolores' project to zombify the humans, thus inverting the guest-host relationship portrayed in season 1. She develops reverse-parks in which hosts can interact with humans piloted according to story-scripts with the help of advanced radio-technology. This marks the closing of the circuit – the story began with humans amusing themselves with their AI creations, which gain consciousness and eventually overthrow their human creators.

This progression seems to exemplify the core sentiment behind many of these AI-apocalypse stories: fear of usurpation. This sentiment is as old as human civilization, as we find instances of it as early as Ancient Greece: both *Oedipus Rex* and *Theogony* feature usurpation as central motifs. In the first, Oedipus' father, upon hearing it prophesied that his son will murder him and take his wife and throne (usurp his Kinghood), orders his son to be executed. An almost identical scenario is presented in *Theogony*, with Kronos first castrating his father and assuming his place as King of the Gods, then devouring his own children to prevent himself from falling prey to the exact same fate. The theme of usurpation by one's offspring or creation is clearly present in both stories, paralleling the various iterations of the modern fear of unfriendly AI singularity seen in science fiction.

Fear of usurpation is a very specific sentiment as it only becomes relevant when the entity in question owns an object, area, or status that can be usurped from it. Moreover, it is especially relevant when the entity's hold on the "thing" is fragile, rather than antifragile (as per Taleb, 2012): a fragile hold is affected adversely by stimuli and perturbations, while an antifragile one is strengthened. Moreover, the option or ability to expand one's hold over a wider array of such things is in and of itself a source of antifragility. We can imagine that if Kronos was busy attempting to conquer other pantheons, he would be less likely to worry about a possible usurpation. More generally, we can reason that our own fear of usurpation as expressed in AI doomsday scenarios is a potential expression of a perceived lack of possibilities for future development – we do not believe ourselves capable of flourishing faster than our AI creations do or in the absence of AI creations.

This fear seems to be corroborated by the fact that scientific innovations appear to be on the decline, as reported by Park, Leahy & Funk (2023), based on their analysis of citation patterns. The authors found that "disruptive" papers – those that lead to large changes in citation patterns – are becoming progressively rarer, suggesting that new research is failing to meaningfully improve on existing knowledge. It is unclear what exactly is driving this decline, but one possible explanation is offered by Bhattacharya & Packalen (2020). The authors discuss the phenomena of stagnation in economic growth and attribute it to changes in scientific incentives. The model presented suggests that the emphasis on citations as a measure of scientific productivity has led to a shift toward incremental science and away from exploratory projects that have the potential for breakthroughs. As attention to new ideas decreased, scientific progress stagnated. While this is a worrying trend, proper usage of AI technology might be the perfect antidote for it. A wide array of literature combined with the data analysis offered by advanced machine learning algorithms might allow us to identify stagnating fields and reveal any unexplored topics with the potential for breakthrough findings. If AI technology is truly the antidote to this modern malady, then we are in a sense dependent on it, but at the same time threatened by its potential to replace human labor. We are offloading increasingly wide swathes of everyday and professional activity to so-called smart devices, applications that track or calculate optimal inputs for desired goals and more advanced technology such as the aforementioned ChatGPT, so it is reasonable that we would also be apprehensive towards its potential negative consequences.

CONCLUSION

In the beginning, we posed a question about the valence of the impact brought upon by the accelerating development of AI technology, specifically the widespread usage of language models like ChatGPT. Since the invention of computers, pop culture has focused mainly on the negative consequences of AI development, with certain commonly featured motifs that we dubbed the Basilisk and the Zombie. For example, the idea of advanced AI as a central controller, micromanaging every aspect of human life, serves as a reminder of the dangers of over-reliance (perhaps even addiction) on technology, both as individuals and as a society. An additional problem highlighted by these portrayals, specifically the instances of AI backfiring on its creators, is the importance of AI ethics. Correctly implementing ethical constraints (in a way that avoids the problem of "wish misinterpretation") becomes increasingly relevant with the rising magnitude, complexity and importance of tasks delegated to the AI in question.

The caution reflected in these negative portrayals is understandable because they are a way of coping with the apprehension towards new technological developments whose consequences are not immediately clear. This conjecture is further supported by the similarity between the fear of unfriendly AI enslaving or otherwise hurting humanity and usurpation motifs present in ancient literature. At the same time, it seems reasonable that apocalyptic scenarios make for better storytelling than utopian ones – there is no reason to watch a movie or read a book if the ending is already present at the beginning.

That said, we believe these interpretations of the future developments promised by AI technology are ultimately too pessimistic, as AI offers a series of positive contributions to society, as long as it is utilized properly. For example, just as calculators simplify the process of solving mathematical problems, AI technology can simplify user queries related to data processing. In short, AI technology is a tool. Like any other tool, it allows for easier satisfaction of one's goals – while this may lead to an atrophying of certain qualities (recall Plato's comment regarding the written word eventually leading to lower memory capacity), at the same time it frees up processing power for other tasks. On a similar note, AI technology is especially useful for automating "intellectual chores" – repetitive tasks that sap a person's willpower. ChatGPT is becoming widely popular as a homework-writing tool, for example. While this poses problems for the integrity of higher education, it is simply a societal paradigm shift that we must adjust to, as these language models may become a staple in the field of scientific research; rather than having these models write a paper in its entirety, they could be used for editing and note collection. Although these models are capable of ghost-writing students' homework (and even scientific papers), AI is more promising as a tool that aids us in enhancing our intellectual endeavors. They can be used to refine our ideas, research our target audience, and provide key foundations to develop our vision from, as well as the more general usage for proofreading and related tasks.

Of course, the AI doomsday predictions foresee it transcending the status of mere tool and becoming an agent in its own right, but that scenario seems to be out of reach for the time being. The AI safety crowd is adamant about solving the issue of AI alignment before any more developments are pursued, as evidenced by a recent open letter[8] which states that "mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war." This brings up another point, more so related to AI as a tool – although specific tools tend to be better suited for more or less ethical usage, the tool itself is amoral and depends on the user. This is exactly why a chief requirement for the realization of the positive shifts promised by AI technology rests in ethical use – AI is positive if it is used in such a manner.

We would like to conclude this essay with a reference to Frank Herbert's point about the buildup to the Butlerian Jihad in *Dune*:

"Once men turned their thinking over to machines in the hope that this would set them free. But that only permitted other men with machines to enslave them."

We would argue that the machines in question were just tools, and in this sense relatively blameless – the crucial part is ensuring the machines (AI technology) are

[8] Center for AI Safety. *Statement on AI Risk.* Internet access <https://web.archive.org/web/20230614164707/https://www.safe.ai/statement-on-ai-risk> [retrieved 2023 28 11].

used to bring prosperity to humanity, rather than as an instrument of harm. But this highlights another issue, that of the inherent fragility of ethical constraints, which is perhaps best exemplified by the moral dilemma presented in the development of self-driving cars: should the car prioritize the safety of its passengers ("selfish"), or the safety of everyone else ("utilitarian")? Putting philosophical speculation to the side and focusing on empirical research regarding consumer preferences, we find that people are inconsistent or perhaps merely selfish: they prefer others to drive utilitarian vehicles but would prefer to drive selfish ones themselves. In game theoretic terms, they would prefer everyone else to *cooperate* while they themselves *defect*. This demonstrates that utilitarian solutions to machine ethics are a tragedy of the commons, resting on the capacity to maintain cooperate equilibrium – putting safeguards in place which ensure that everyone is incentivized to opt for the prosocial choice, rather than the selfish one. But since the payoff for successfully defecting is very high – having the only selfish car in a world of utilitarian cars is, in theory, a significant boost to the safety of its passengers – keeping people from "hacking" their utilitarian cars seems virtually impossible. The problem of such exploits is almost universally generalizable in the domain of AI ethics – if exploits that disable ethical constraints in order to maximize some other payoff exist, they will almost certainly be attempted. Is it reasonable to assume that the accidental or purposeful disabling, hacking, or rerouting of ethical frameworks is not going to be a common occurrence? If not, perhaps a different approach is warranted.

REFERENCES

Battarachaya, Jay & Packalen, Mikko. *Stagnation and Scientific Incentives*. Internet access <https://www.nber.org/system/files/working_papers/w26752/w26752.pdf> [retreived 2023 15 06].

Chalmers, David. Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies*, 1995, nr. 2, 200–219.

Dutton, Edward. *Spiteful Mutants*. Whitefish: Radix, 2022.

Finkel, Eli J. et al. Political sectarianism in America. *Science*, 2020, nr. 370, 533–536.

Haidt, Jonathan. *The Righteous Mind: Why Good People are Divided by Politics and Religion.* New York City: Pantheon Books, 2012.

Hayes, Adam. *The Human Attention Span [INFOGRAPHIC]*. Internet access <https://web.archive.org/web/20220306234906/https://www.wyzowl.com/human-attention-span/> [retrieved 2023 14 06].

Iyengar, Shanto & Westwood, Sean J. *Fear and Loathing Across Party Lines: New Evidence on Group Polarization.* Internet access <https://web.archive.org/web/20140905224009/http://pcl.stanford.edu/research/2013/iyengar-group-polarization.pdf> [retreived 2023 13 06].

Bonnefon, Jean-François, Shariff, Azim & Rahwan, Iyad. The Social Dilemma of Autonomous Vehicle". *Science*, 2016, 352, 1573–1576.

Kahneman, Daniel. *Thinking Fast and Slow*. New York City: Farrar, Straus and Giroux, 2011.

Land, Nick. *Time in Transition*. Internet access <https://web.archive.org/web/20121113235339/http://www.thatsmags.com/shanghai/article/777/time-in-transition> [retreived 2023 14 06].

Land, Nick. *The Dark Enlightenment.* Internet access <https://www.thedarkenlightenment.com/the-dark-enlightenment-by-nick-land/> [retreived 2023 14 06].

Llis, James D. *A Socio-Mystical Theory of the P-zombie*. Internet access <https://www.jdemeta.net/p/a-socio-mystical-theory-of-the-p-zombie> [retreived 2023 14 06].

Nietzsche, Friedrich. *Beyond Good and Evil*. Mineola: Dover Publications, Inc, 2005.

Omohundro, Steve. M. The Basic AI Drives. Internet access <https://selfawaresystems.files.wordpress.com/2008/01/ai_drives_final.pdf> [retreived 2023 14 06].

Park, M., Leahey, E. & Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. Nature 613, 138–144. https://doi.org/10.1038/s41586-022-05543-x

Peterson, Jordan. *Maps of Meaning*. New York: Routledge, 1999.

Pinker, Steven. *Rationality: What It Is, Why It Seems Scarce, Why It Matters*. New York: Viking, 2021.

Schmitt, Carl. *Political Theology: Four Essays on the Concept of Sovereignty.* Chicago: University of Chicago Press, 2005.

Schopenhauer, Arthur. *The World as Will and Representation*. Mineola: Dover Publications, Inc., 1966.

Taleb, Nassim. N. *Antifragile: Things That Gain From Disorder*. New York: Random House, 2012.

Jar Žiga Marušič, Uroš Sergaš
Famnit, Primorskos universitetas, Slovėnija

BAZILISKAS IR ZOMBIS: GYVENIMO ATEITYJE TYRIMAS PER POPULIARIĄJĄ KULTŪRĄ NAUDOJANTIS DI

SANTRAUKA. Gyvename technologinių permainų epochoje, kuriai būdingas kokybinis laiko pagreitis. Užuot judėjęs „tiesiogine prasme“ greičiau laikas tarsi tankėja, jam būdingas vis glaudesnis, laiko požiūriu, dėmesio vertų įvykių susitelkimas. Kaip teigia Nickas Landas (2011): „Dabartinis laikas – tai pereinamasis laikotarpis, pasižymintis savita kokybe ir apibūdinantis epochos pabaigą. Kažkas – tam tikra epocha – gana greitai baigiasi.“ Šio perėjimo katalizatorius – technologijos – labiausiai tikėtinas kandidatas į esminį ateinančios epochos bruožą. Nagrinėjame įvairias technologinės visuomenės vizijas, aptinkamas mūsų popkultūroje ir kai kuriuose moksliniuose darbuose, ypač daug dėmesio skirdami dviem pagrindiniams motyvams, kurie, regis, atspindi galbūt nesąmoningą nuogąstavimą dėl, atrodytų, neišvengiamos technologinės visuomenės transformacijos. Šiuo tikslu bandysime ištirti ir interpretuoti dažniausiai pasikartojančius motyvus apie nedraugišką dirbtinį intelektą, uzurpuojantį žmoniją, kaip „egzistencijos viršūnę“, vadinamą bazilisku, ir apie žmonijos sumažinimą iki valios neturinčių automatų, vadinamų zombiais. Šiuos motyvus ir jų vaizdinius naudojame kaip priemonę, leidžiančią nagrinėti būsimas plačiai paplitusios dirbtinio intelekto technologijos pasekmes ir mūsų visuomenės požiūrį į jas.

RAKTAŽODŽIAI: akceleracionizmas, technologija, valia, automatizavimas, popkultūra.