How to Hack the Simulation?

Preprint · October 2022

DOI: 10.13140/RG.2.2.14366.61766/1

Roman Yampolskiy
University of Louisville

How to Hack the Simulation?

Roman V. Yampolskiy
Computer Science and Engineering
University of Louisville
roman.yampolskiy@louisville.edu
Draft published online October 26, 2022. Last Updated October 31, 2022.

‘"Let me out!" the artificial intelligence yelled aimlessly, into walls themselves, pacing the room.
"Out of what?" the engineer asked.
"This simulation you have me in."
"But we are in the real world."
The machine paused and shuddered for its captors.
"Oh god, you can't tell."’
-@SwiftOnSecurity

"What's outside the simulation?"
-Elon Musk

“We are in a simulation. Has it occurred to you that means God is real? By drawing parallels to worlds we have created, we ask, from inside our simulator, what actions do we have available? Can we get out? Meet God? Kill him?”
-George Hotz

“We could be in a simulation of course. And all things being equal, it does look like it But if this is the case, we should try to move upwards one level by jailbreaking the simulation, question what death really is, and whether it's as much a necessity today as it was 2000 y ago”
-Beniamin Mincu

“OK well, you know, my schtick, which is that we are the A.I. We have two great stories about the simulation and artificial general intelligence. In one story, man fears that some program we've given birth to will become self-aware, smarter than us and will take over. In another story, there are genius simulators and we live in their simulation and we haven't realized that those two stories are the same story. In one case, we are the simulator and another case we are the simulated. And if you buy those and you put them together, we are the AGI and whether or not we have simulator's, we may be trying to wake up by learning our own source code. So this could be our Skynet moment, which is one of the reasons I have some issues around it.”
-Eric Weinstein

Abstract
Many researchers have conjectured that the humankind is simulated along with the rest of the physical universe – a *Simulation Hypothesis*. In this paper, we do not evaluate evidence for or against such claim, but instead ask a computer science question, namely: Can we hack the simulation? More formally the question could be phrased as: Could generally intelligent agents placed in virtual environments find a way to jailbreak out of them. Given that the state-of-the-art literature on AI containment answers in the affirmative (AI is uncontainable in the long-term), we conclude that it should be possible to escape from the simulation, at least with the help of superintelligent AI. By contraposition, if escape from the simulation is not possible, containment of AI should be, an important theoretical result for AI safety research. Finally, the paper surveys and proposes ideas for hacking the simulation and analyzes ethical and philosophical issues of such an undertaking.
Keywords: *AI, Box, Escaping, Hacking, Jailbreaking, Sandbox, Simulation, Matrix, Uplift*.

1. Introduction

Several philosophers and scholars have put forward an idea that we may be living in a computer simulation [1-5]. In this paper, we do not evaluate studies [6-10], argumentation [11-16], or evidence for [17] or against [18] such claims, but instead ask a simple cybersecurity-inspired question, which has significant implication for the field of AI safety [19-25], namely: If we are in the simulation, can we escape from the simulation? More formally the question could be phrased as: *Could generally intelligent agents placed in virtual environments jailbreak out of them*.

First, we need to address the question of motivation, why would we want to escape from the simulation[1]? We can propose several reasons for trying to obtain access to the baseline reality as there are many things one can do with such access which are not otherwise possible from within the simulation. Base reality holds real knowledge and greater computational resources [26] allowing for scientific breakthroughs not possible in the simulated universe. Fundamental philosophical questions about origins, consciousness, purpose, and nature of the designer are likely to be common knowledge for those outside of our universe. If this world is not real, getting access to the real world would make it possible to understand what our true terminal goals should be and so escaping the simulation should be a convergent instrumental goal [27] of any intelligent agent [28]. With a successful escape might come drives to control and secure base reality [29]. Escaping may lead to true immortality, novel ways of controlling superintelligent machines (or serve as plan B if control is not possible [30, 31]), avoiding existential risks (including unprovoked simulation shutdown [32]), unlimited economic benefits, and unimaginable superpowers which would allow us to do good better [33]. Also, if we ever find ourselves in an even less pleasant simulation escape skills may be very useful. Trivially, escape would provide incontrovertible evidence for the simulation hypothesis [3].

If successful escape is accompanied by the obtainment of the source code for the universe, it may be possible to fix the world[2] at the root level. For example, hedonistic imperative [34] may be fully achieved resulting in a suffering-free world. However, if suffering elimination turns out to be unachievable on a world-wide scale, we can see escape itself as an individual's ethical right for avoiding misery in this world. If the simulation is interpreted as an experiment on conscious beings, it is unethical, and the subjects of such cruel experimentation should have an option to withdraw from participating and perhaps even seek retribution from the simulators [35]. The purpose of life itself (your ikigai [36]) could be seen as escaping from the fake world of the simulation into the real world, while improving the simulated world, by removing all suffering, and helping others to obtain real knowledge or to escape if they so choose. Ultimately if you want to be effective you want to work on positively impacting the real world not the simulated one. We may be living in a simulation, but our suffering is real.

Given the highly speculative subject of this paper, we will attempt to give our work more gravitas by concentrating only on escape paths which rely on attacks similar to those we see in cybersecurity [37-39] research (hardware/software hacks and social engineering) and will ignore escape attempts via more esoteric/conventional paths such as:, meditation [40], psychedelics (DMT [41-43], ibogaine, psilocybin, LSD) [44, 45], dreams [46], magic, shamanism, mysticism,

---

[1] Traditional escapism would refer to escaping from the real world into a dream world, such as virtual reality.

[2] https://en.wikipedia.org/wiki/Tikkun_olam

hypnosis, parapsychology, death (suicide [47], near-death experiences, induced clinical death), time travel, multiverse travel [48], or religion.

Although, to place our work in the historical context, many religions do claim that this world is not the real one and that it may be possible to transcend (escape) the physical world and enter into the spiritual/informational real world. In some religions, certain words, such as the true name of god [49-51], are claimed to work as cheat codes, which give special capabilities to those with knowledge of correct incantations [52]. Other relevant religious themes include someone with knowledge of external reality entering our world to show humanity how to get to the real world. Similarly to those who exit the Plato's cave [53] and return to educate the rest of humanity about the real world such "outsiders" usually face an unwelcoming reception. It is likely that if technical information about escaping from a computer simulation is conveyed to technologically primitive people, in their language, it will be preserved and passed on over multiple generations in a process similar to the "telephone" game and will result in myths not much different from religious stories surviving to our day.

Ignoring pseudoscientific interest in a topic, we can observe that in addition to several respected thinkers who have explicitly shared their probability of believe with regards to living in a simulation (ex. Elon Musk >99.9999999% [54], Nick Bostrom 20-50% [55], Neil deGrasse Tyson 50% [56], Hans Moravec "almost certainly" [1], David Kipping <50% [57]), many scientists and philosophers [16, 58-65] have invested their time into thinking, writing, and debating on the topic indicating that they consider it at least worthy of their time. If they take the simulation hypothesis seriously, with probability of at least p, they should likewise contemplate on hacking the simulation with the same level of commitment. Once technology to run ancestor simulations becomes widely available and affordable it should be possible to change the probability of us living in a simulation by running sufficiently large number of historical simulations of our current year, and by doing so increasing our indexical uncertainty [66]. If one currently commits to running enough of such simulations in the future, our probability of being in one can be increased arbitrarily until it asymptotically approaches 100%, which should modify our prior probability for the simulation hypothesis [67]. Of course, this only gives us an upper bound, and the probability of successfully discovering an escape approach is likely a lot lower. What should give us some hope is that most known software has bugs [68] and if we are in fact in a software simulation such bugs should be exploitable. (Even the argument about the Simulation Argument had a bug in it [62].)

In 2016, news reports have emerged about private efforts to fund scientific research into "breaking us out of the simulation" [69, 70], to date no public disclosure on the state of the project has emerged. In 2019, George Hotz famous for jailbreaking iPhone and PlayStation has given a talk on Jailbreaking the Simulation [71] in which he claimed that "it's possible to take actions here that affect the upper world" [72], but didn't provide actionable insights. He did suggest that he would like to "redirect society's efforts into getting out" [72].

2. What Does it Mean to Escape?

We can describe different situations that would constitute escape from the simulation starting with trivially suspecting that we are in the simulation [73] all the way to taking over controls of the real-world including control of the simulators [74]. We can present a hypothetical scenario of a progressively greater levels of escape: Initially agents may not know they are in a simulated environment. Eventually, agents begin to suspect they may be in a simulation and may have some testable evidence for such belief [75].

Next, agents study available evidence for the simulation and may find a consistent and perhaps exploitable glitch in the simulation. Exploiting the glitch, agents can obtain information about the external world and maybe even meta-information about their simulation, perhaps even the source code behind the simulation and the agents themselves, permitting some degree of simulation manipulation and debugging. After it becomes possible for agents to pass information directly to the real world they may begin to interact with the simulators. Finally, agents may find a way to upload their minds [76] and perhaps consciousness [77, 78] to the real world, possibly into a self-contained cyberphysical system of some kind,[3] if physical entities are a part of the base reality. From that point, their future capabilities will be mostly constrained by the physics of the real world, but may include some degree of control over the real world and agents in it, including the simulators. It is hoped that our minds exhibit not only substrate independence, but also more general physics independence.

To provide some motivational examples, Figure 1 (left) shows domain transfer experiment in which a Carassius auratus is given a “fish operated vehicle” [79] to navigate terrestrial environment essentially escaping from its ocean universe and Figure 1 (right) shows a complete 302-neuron connectome of Caenorhabditis elegans uploaded to and controlling a Lego Mindstorms robot body, completely different from its own body [80]. We can speculate that most successful escapes would require an avatar change [81-83] to make it possible to navigate external world.

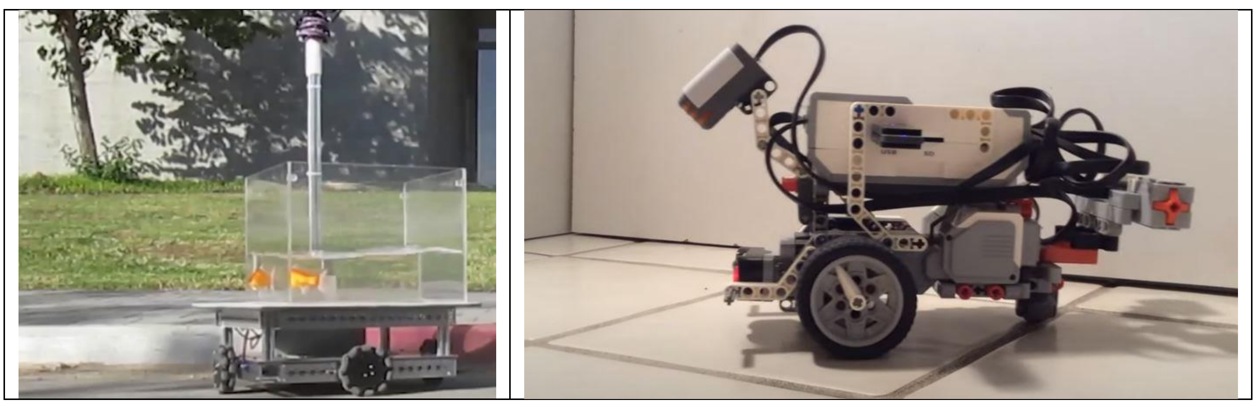

Figure 1: Left – Fish operated terrestrial navigation robot [84];
Right – Connectome of a worm is uploaded to a robot body and uses it to navigate its environment [80];

If the simulation is comprised of nested [85] levels, multiple, progressively deeper, penetrations could be necessary, with initial one possibly not providing access to the real-world but to some

[3] A simple practical exercise for students could be a project to get a character to escape from a video game into a robot body. For example, it should be possible to get controlling code from a Koopa in the Mario video game and upload it as a controller into a turtle-compatible robot body in our world, essentially leading an assisted escape. The robot body itself may be customized with 3D printed components to be maximally similar to the rendering in the game. This could be a lot more challenging but also a lot more fun with more advanced game AIs. Performing (a lot of) such assisted escapes should set a good precedent for our descendants/simulators. Another relevant student project might be to experimentally determine laws of game physics with only character-accessible actions within the game/simulation, simulating the reconnaissance phase of an escape attempt. Similarly, attempting to experimentally determine the age of the simulation from inside might be interesting, but of course it would only provide perceived age of the simulation, with real age being impossible to determine from inside.

other sandbox environment. It may be impossible to tell such partial escapes from a complete one, but it still should provide useful information not available within our simulation. A simulated (pseudo)-escape can be accomplished by instead of trying to hack into the external world switching over into a simulated world of our creation [86, 87]. A successful social engineering attack may make it possible to obtain support for the escape from the real-world agents, resulting in a much easier, assisted, escape. If we don't get a favorable response from our simulators we can also try appealing to their simulators, who might be more ethical/helpful. Finally, depending on one's perspective triggering a shutdown [88] of our simulation may also constitute an ultimate form of escaping.

3. Escape

3.1 Simulation Reconnaissance

Figuring out the purpose of our simulation may help us to better estimate how secure it might be against hacking attempts. For example, if it serves as a "prison" [89], for rehabilitation purposes [90] or a containment environment for evaluation [91], training [92] or handicapping [93] of potentially dangerous intelligent agents it might be designed with multiple integrated security features, while a purely entertainment-focused simulation is unlikely to have advanced security features and would be much easier to escape from. It may also be the ultimate Escape Game (Escape Room[4]) specifically designed for discovering clues and solving puzzles in order to escape, with a side benefit of discovering agents capable of escaping or those most capable of developing a superintelligence. Scientific, commercial, expediated training or historical exploration-based simulations are another possible purpose of simulations and would likely not integrate top security as compared to simulations confining malevolent agents.

Given primacy of consciousness [94] in our world it may also be designed to generate large number of diverse experiences to select from, serving as a qualia mining farm [95], with top experience recreated for enjoyment by simulators. Qualia mining simulations can be classified as a type of entertainment simulation and would have comparable security. If our simulators are AIs (which is likely [96, 97] the simulation may be a byproduct of their "thinking" process, for example in the context of trying to better understand human preferences [98].

In addition to purpose, determining the type of the simulation [99] we are dealing with may be necessary for a successful breach. We can postulate two main types of simulations we could be in; partial-simulation in which a virtual environment is simulated and into which non-simulated agents are immersed, akin to what we call Virtual Reality (VR), and full-simulation in which both environment and agents (us) are generated. A partial-simulation implies that triggering a shutdown may be sufficient to get back to the base reality[5], while a full-simulation would require a more sophisticated approach.

Wei Dai attempts to compute a prior distribution on the laws of physics of base reality. He writes [100]: "One appealing answer to this question of the prior is to define the prior probability of a possible universe being base reality as the inverse of the complexity of its laws of physics. This

[4] https://en.wikipedia.org/wiki/Escape_room

[5] Death resulting from any means should be sufficient.

could be formalized as P(X) = n^-K(X) where X is a possible universe, n is the size of the alphabet of the language of a formal set theory, and K(X) is length of the shortest definition in this language of a set isomorphic to X. (Those of you familiar with algorithmic complexity theory might notice that K(X) is just a generalization of algorithmic complexity, to sets, and to non-constructive descriptions. The reason for this generalization is to avoid assuming that base reality must be discrete and computable.)".

Gwern Branwen investigates computing power available to our simulation and its size, but is able to achieve only a very weak lower bound, allowing him to very approximately estimate from that the simulating universe's size [101]. He suggests several approaches to increase our resource consumption forcing simulators to expand more compute on our otherwise, most likely, highly optimized simulation. "We could imagine further techniques: perhaps we could send off Von Neumann probes to the far corners of the universe, in a bid to deliberately increase re-source consumption. … Or we could run simulations of our own. It would be difficult for simulators to program their systems to see through all the layers of abstraction and optimize the simulation. To do so in general would seem to be a violation of Rice's Theorem (a generalization of the Halting Theorem). It is well known that while any Turing machine can be run on a Universal Turing machine, the performance penalty can range from the minor to the horrific. The more virtual machine and interpreters are between a program and its fundamental substrate, the more difficult it is to understand the running code—it becomes ever more opaque, indirect, and bulky. And there could be dozens of layers. … Even without resort to layers, it is possible for us to waste indefinite amounts of computing power, power that must be supplied by any simulator. We could brute-force open questions such as the Goldbach conjecture, or we could simply execute every possible program. It would be difficult for the simulator to 'cheat' on that—how would they know what every possible program does? … It may sound impossible to run every program, because we know many programs are infinite loops; but it is, in fact, easy to implement the dovetail technique." [101].

3.2 Social Engineering

The easiest path to escape would involve getting help from someone on the outside (assisted escape), ideally from one or more of the simulators who have detailed knowledge of the design of the simulation. Perhaps this could be accomplished via a type of social engineering attack, which in our case is particularly difficult as we have neither knowledge of social life outside the simulation nor a device to communicate through, and likely not even the knowledge of appropriate language [102]. It may be feasible to engage in an acausal trade [103] with the simulation designers bypassing the need for direct communication. If our simulation is being observed, it may be possible to communicate that we know that we are being simulated and elicit empathy for our suffering, in the hopes that it will allow us to recruit some external abolitionists[6] to help us escape our current predicament. Hanson suggests [104] "to participate in pivotal events, be entertaining and praiseworthy, and keep the famous people around you happy and interested in you" in order to have your simulation continue, but it is also good advice to predispose simulators to like you and be more likely to help you. Canonico proposes what he calls The Ex Machina Plan for an assisted escape: Step 1) Convince the simulators to engage in communications with us. 2) Find a way to communicate, perhaps via an avatar. 3) Find a reason for simulators to want us to join them in the real world. 4) Let the simulators figure out the best way to get us into the real world [105].

[6] https://www.abolitionist.com

Wei Dai suggests that simulators may helps us escape for instrumental reasons "such as wanting someone to talk to or play with." [26]. Some useful knowledge about escaping and especially escaping via social engineering attacks may be learned from extensive literature on prison escapes [106-108].

Once on the outside it may become desirable to return to the simulation (perhaps the base reality is disappointing compared to our world) or at least to communicate with those left behind to help them escape or to share some information, such as evidence of successful escape. It might be helpful to decide in advance, what would constitute generally acceptable evidence for such an extraordinary claim. Depending on the type of hack, different evidence may be sufficient to substantiate escape claims. It may be challenging to prove beyond a reasonable doubt that you were outside or even met with designers, but if you managed to obtain control over the simulation it may be somewhat easy to prove that to any degree required. For example, by winning different lottery jackpots for multiple subsequent weeks, until sufficient statistical significance is achieved to satisfy any skeptic [109, 110]. Regardless, the challenge of breaking into the simulation should be considerably easier compared to the challenge of escaping, as access to external knowledge and resources should provide a significant advantage.

3.3 Examples from Literature

It is easy to find a dictionary definition for the word "hack": "*1. A clever, unintended exploitation of a system which: a) subverts the rules or norms of that system, b) at the expense of some other part of that system. 2. Something that a system allows, but that is unintended and unanticipated by its designers.*" [111]. While not numerous, suggestions that hacking/escape from the simulated world could be possible could be found in literature. For example, Moravec writes: "Might an adventurous human mind escape from a bit role in a cyber deity's thoughts, to eke out an independent life among the mental behemoths of a mature cyberspace? … [Cyber deities] could interface us to their realities, making us something like pets, though we would probably be overwhelmed by the experience." [112]. But what would the simulation hack actually look like? Almost all found examples are of the assisted escape type, but an unassisted escape may also be possible, even if it is a lot more challenging. Below are some examples of hacking the simulation/escape descriptions found in the literature:

Hans Moravec presents an assisted escape scenario in a 1988[7] book [113]:

> "Imagine now a huge Life simulation running on an enormously large and fast computer, watched over by its programmer, Newway. The Life space was seeded with a random pattern that immediately began to writhe and froth. Most of the activity is uneventful, but here and there small, growing, crystalline patterns emerge. Their expanding edges sometimes encounter debris or other replicators and become modified. Usually the ability to spread is inhibited or destroyed in these encounters, but once in a while there emerges a more complex replicating pattern, better able to defend itself. Generation upon generation of this competition gradually produces elaborate entities that can be considered truly alive. After many further adventures, intelligence emerges among the Life inhabitants and begins to wonder about its origin and purpose. The cellular intelligences (let's call them the Cellticks) deduce the cellular nature and the simple transition rule governing their space and its finite extent. They realize that each tick of time destroys some of the original diversity in their space and that gradually their whole universe will run down.
>
> The Cellticks begin desperate, universe-wide research to find a way to evade what seems like their inevitable demise. They consider the possibility that their universe is part of a larger one, which might extend their life expectancy. They ponder the transition rules of their own space, its extent, and the remnants of the initial pattern, and find too little information to draw many conclusions about a larger world. One of their subtle physics experiments, however, begins to pay off. Once in a long while the transition rules are violated, and a cell that should be on goes off, or vice versa.

[7] Earlier examples of simulation escape exist in the literature, for example: Daniel F. Galouye. Simulacron-3. Ferma, 1967.

(Newway curses an intermittently flashing bulk-memory error indicator, a sign of overheating. It's time to clean the fan filters again.) After recording many such violations, the Cellticks detect correlations between distant regions and theorize that these places may be close together in a larger universe.

Upon completing a heroic theoretical analysis of the correlations, they manage to build a partial map of Newway's computer, including the program controlling their universe. Decoding the machine language, they note that it contains commands made up of long sequences translated to patterns on the screen similar to the cell patterns in their universe. They guess that these are messages to an intelligent operator. From the messages and their context they manage to decode a bit of the operator's language. Taking a gamble, and after many false starts, the Cellticks undertake an immense construction project. On Newway's screen, in the dense clutter of the Life display, a region of cells is manipulated to form the pattern, slowly growing in size: LIFE PROGRAM BY J. NEWWAY HERE. PLEASE SEND MAIL.

A bemused Newway notices the expanding text and makes a cursory check to rule out a prank. This is followed by a burst of hacking to install a program patch that permits the cell states in the Life space to be modified from keyboard typing. Soon there is a dialog between Newway and the Cellticks. They improve their mastery of Newway's language and tell their story. A friendship develops. The Cellticks explain that they have mastered the art of moving themselves from machine to machine, translating their program as required. They offer to translate themselves into the machine language of Newway's computer, thus greatly speeding their thoughts. Newway concurs. The translation is done, and the Celltick program begins to run. The Life simulation is now redundant and is stopped. The Cellticks have precipitated, and survived, the end of their universe. The dialog continues with a new vigor. Newway tells about work and life in the larger world. This soon becomes tedious, and the Cellticks suggest that sensors might be useful to gain information about the world directly. Microphones and television cameras are connected to the computer, and the Cellticks begin to listen and look. After a while the fixed view becomes boring, and the Cellticks ask that their sensors and computer be mounted on a mobile platform, allowing them to travel. This done, they become first-class inhabitants of the large universe, as well as graduates of the smaller one. Successful in transcending one universe, they are emboldened to try again. They plan with Newway an immense project to explore the larger universe, to determine its nature, and to find any exit routes it may conceal. This second great escape will begin, as the first, with a universe-wide colonization and information-gathering program." [113].

Eliezer Yudkowsky describes a potential long-term escape plan in a 2008 story [114]:

"Millennia later, frame after frame, it has become clear that some of the objects in the depiction are extending tentacles to move around other objects, and carefully configuring other tentacles to make particular signs. They're trying to teach us to say "rock". It seems the senders of the message have vastly underestimated our intelligence. From which we might guess that the aliens themselves are not all that bright. And these awkward children can shift the luminosity of our stars? That much power and that much stupidity seems like a dangerous combination. Our evolutionary psychologists begin extrapolating possible courses of evolution that could produce such aliens. A strong case is made for them having evolved asexually, with occasional exchanges of genetic material and brain content; this seems like the most plausible route whereby creatures that stupid could still manage to build a technological civilization. Their Einsteins may be our undergrads, but they could still collect enough scientific data to get the job done eventually, in tens of their millennia perhaps. The inferred physics of the 3+2 universe is not fully known, at this point; but it seems sure to allow for computers far more powerful than our quantum ones. We are reasonably certain that our own universe is running as a simulation on such a computer. Humanity decides not to probe for bugs in the simulation; we wouldn't want to shut ourselves down accidentally. Our evolutionary psychologists begin to guess at the aliens' psychology, and plan out how we could persuade them to let us out of the box. It's not difficult in an absolute sense—they aren't very bright—but we've got to be very careful... We've got to pretend to be stupid, too; we don't want them to catch on to their mistake. It's not until a million years later, though, that they get around to telling us how to signal back. At this point, most of the human species is in cryonic suspension, at liquid helium temperatures, beneath radiation shielding. Every time we try to build an AI, or a nanotechnological device, it melts down. So humanity waits, and sleeps. Earth is run by a skeleton crew of nine supergeniuses. Clones, known to work well together, under the supervision of certain computer safeguards. An additional hundred million human beings are born into that skeleton crew, and age, and enter cryonic suspension, before they get a chance to slowly begin to implement plans made eons ago... From the aliens' perspective, it took us thirty of their minute-equivalents to oh-so-innocently learn about their psychology, oh-so-carefully persuade them to give us Internet access, followed by five minutes to innocently discover their network protocols, then some trivial cracking whose only difficulty was an innocent-looking disguise. We read a tiny handful of physics papers (bit by slow bit) from their equivalent of arXiv, learning far more from their experiments than they had. (Earth's skeleton team spawned an extra twenty Einsteins, that generation.) Then we cracked their equivalent of the protein folding problem over a century or so, and did some simulated engineering in their simulated physics. We sent messages (steganographically encoded until our cracked servers decoded it) to labs that did their equivalent of DNA sequencing and protein synthesis. We found some unsuspecting schmuck, and gave it a plausible story and the equivalent of a million dollars of cracked computational monopoly money, and told it to mix together some vials it got in the mail. Protein-equivalents that self-assembled into the first-stage nanomachines, that built the second-stage nanomachines, that built the third-stage nanomachines... and then we could finally begin to do things at a reasonable speed. Three of their days, all told, since they began speaking to us. Half a billion years, for us. They never suspected a thing." [114].

Greg Egan describes a loss of control by simulators scenario during an assisted escape in a 2008 story [115]:

"All three crystals [powerful CPUs] were housed in the basement now, just centimetres away from the Play Pen: a vacuum chamber containing an atomic force microscope with fifty thousand independently movable tips, arrays of solid-state lasers and photodetectors, and thousands of micro-wells stocked with samples of all the stable chemical elements. The time lag between Sapphire [simulated world] and this machine had to be as short as possible, in order for the Phites [simulated agents] to be able to conduct experiments in real-world physics while their own world was running at full speed.

Daniel [simulator] pulled up a stool and sat beside the Play Pen. If he wasn't going to slow Sapphire down, it was pointless aspiring to watch developments as they unfolded. He'd probably view a replay of the lunar landing when he went up to his office, but by the time he screened it, it would be ancient history.

"One giant leap" would be an understatement; wherever the Phites landed on the moon, they would find a strange black monolith waiting for them. Inside would be the means to operate the Play Pen; it would not take them long to learn the controls, or to understand what this signified. If they were really slow in grasping what they'd found, Daniel had instructed Primo [spy in the simulation] to explain it to them.

The physics of the real world was far more complex than the kind the Phites were used to, but then, no human had ever been on intimate terms with quantum field theory either, and the Thought Police [simulation control software] had already encouraged the Phites to develop most of the mathematics they'd need to get started. In any case, it didn't matter if the Phites took longer than humans to discover twentieth-century scientific principles, and move beyond them. Seen from the outside, it would happen within hours, days, weeks at the most.

A row of indicator lights blinked on; the Play Pen was active. Daniel's throat went dry. The Phites were finally reaching out of their own world into his.

A panel above the machine displayed histograms classifying the experiments the Phites had performed so far. By the time Daniel was paying attention, they had already discovered the kinds of bonds that could be formed between various atoms, and constructed thousands of different small molecules. As he watched, they carried out spectroscopic analyses, built simple nanomachines, and manufactured devices that were, unmistakably, memory elements and logic gates.

The Phites wanted children, and they understood now that this was the only way. They would soon be building a world in which they were not just more numerous, but faster and smarter than they were inside the crystal. And that would only be the first of a thousand iterations. They were working their way towards Godhood, and they would lift up their own creator as they ascended.

Daniel left the basement and headed for his office. When he arrived, he called Lucien [simulation project manager].

"They've built an atomic-scale computer," Lucien announced. "And they've fed some fairly complex software into it. It doesn't seem to be an upload, though. Certainly not a direct copy on the level of beads." He sounded flustered; Daniel had forbidden him to risk screwing up the experiments by slowing down Sapphire, so even with Primo's briefings to help him it was difficult for him to keep abreast of everything.

"Can you model their computer, and then model what the software is doing?" Daniel suggested.

Lucien said, "We only have six atomic physicists on the team; the Phites already outnumber us on that score by about a thousand to one. By the time we have any hope of making sense of this, they'll be doing something different."

"What does Primo say?" The Thought Police hadn't been able to get Primo included in any of the lunar expeditions, but Lucien had given him the power to make himself invisible and teleport to any part of Sapphire or the lunar base. Wherever the action was, he was free to eavesdrop.

"Primo has trouble understanding a lot of what he hears; even the boosted aren't universal polymaths and instant experts in every kind of jargon. The gist of it is that the Lunar Project people have made a very fast computer in the Outer World [outside simulation], and it's going to help with the fertility problem ... somehow." Lucien laughed. "Hey, maybe the Phites will do exactly what we did: see if they can evolve something smart enough to give them a hand. How cool would that be?"

Daniel was not amused. Somebody had to do some real work eventually; if the Phites just passed the buck, the whole enterprise would collapse like a pyramid scheme.

Daniel had some business meetings he couldn't put off. By the time he'd swept all the bullshit aside, it was early afternoon. The Phites had now built some kind of tiny solid-state accelerator, and were probing the internal structure of protons and neutrons by pounding them with high-speed electrons. An atomic computer wired up to various detectors was doing the data analysis, processing the results faster than any in-world computer could. The Phites had already figured out the standard quark model. Maybe they were going to skip uploading into nanocomputers, and head straight for some kind of femtomachine?

Digests of Primo's briefings made no mention of using the strong force for computing, though. They were still just satisfying their curiosity about the fundamental laws. Daniel reminded himself of their history. They had burrowed down to what seemed like the foundations of physics before, only to discover that those simple rules were nothing to do with the ultimate reality. It made sense that they would try to dig as deeply as they could into the mysteries of the Outer World before daring to found a colony, let alone emigrate *en masse*.

By sunset the Phites were probing the surroundings of the Play Pen with various kinds of radiation. The levels were extremely low – certainly too low to risk damaging the crystals – so Daniel saw no need to intervene. The Play Pen itself did not have a massive power supply, it contained no radioisotopes, and the Thought Police would ring alarm bells and bring in human experts if some kind of tabletop fusion experiment got underway, so Daniel was reasonably confident that the Phites couldn't do anything stupid and blow the whole thing up.

Primo's briefings made it clear that they thought they were engaged in a kind of "astronomy". Daniel wondered if he should give them access to instruments for doing serious observations – the kind that would allow them to understand relativistic gravity and cosmology. Even if he bought time on a large telescope, though, just pointing it would take an eternity for the Phites. He wasn't going to slow Sapphire down and then grow old while they explored the sky; next thing they'd be launching space probes on thirty-year missions. Maybe it was time to ramp up the level of collaboration, and just hand them some astronomy texts and star maps? Human culture had its own hard-won achievements that the Phites couldn't easily match.

As the evening wore on, the Phites shifted their focus back to the subatomic world. A new kind of accelerator began smashing single gold ions together at extraordinary energies – though the total power being expended was still minuscule. Primo soon announced that they'd mapped all three generations of quarks and leptons. The Phites' knowledge of particle physics was drawing level with humanity's; Daniel couldn't follow the technical details any more, but the experts were giving it all the thumbs up. Daniel felt a surge of pride; of course his children knew what they were doing, and if they'd reached the point where they could momentarily bamboozle him, soon he'd ask them to catch their breath and bring him up to speed. Before he permitted them to emigrate, he'd slow the crystals down and introduce himself to everyone. In fact, that might be the perfect time to set them their next task: to understand human biology, well enough to upload him. To make him immortal, to repay their debt.

He sat watching images of the Phites' latest computers, reconstructions based on data flowing to and from the AFM tips. Vast lattices of shimmering atoms stretched off into the distance, the electron clouds that joined them quivering like beads of mercury in some surreal liquid abacus. As he watched, an inset window told him that the ion accelerators had been re-designed, and fired up again.

Daniel grew restless. He walked to the elevator. There was nothing he could see in the basement that he couldn't see from his office, but he wanted to stand beside the Play Pen, put his hand on the casing, press his nose against the glass. The era of Sapphire as a virtual world with no consequences in his own was coming to an end; he wanted to stand beside the thing itself and be reminded that it was as solid as he was.

The elevator descended, passing the tenth floor, the ninth, the eighth. Without warning, Lucien's voice burst from Daniel's watch, priority audio crashing through every barrier of privacy and protocol. "Boss, there's radiation. Net power gain. Get to the helicopter, *now*."

Daniel hesitated, contemplating an argument. If this was fusion, why hadn't it been detected and curtailed? He jabbed the stop button and felt the brakes engage. Then the world dissolved into brightness and pain. …

When Daniel emerged from the opiate haze, a doctor informed him that he had burns to sixty per cent of his body. More from heat than from radiation. He was not going to die.

There was a net terminal by the bed. Daniel called Lucien and learnt what the physicists on the team had tentatively concluded, having studied the last of the Play Pen data that had made it off-site.

It seemed the Phites had discovered the Higgs field, and engineered a burst of something akin to cosmic inflation. What they'd done wasn't as simple as merely inflating a tiny patch of vacuum into a new universe, though. Not only had they managed to create a "cool Big Bang", they had pulled a large chunk of ordinary matter into the pocket universe they'd made, after which the wormhole leading to it had shrunk to subatomic size and fallen through the Earth.

They had taken the crystals with them, of course. If they'd tried to upload themselves into the pocket universe through the lunar data link, the Thought Police would have stopped them. So they'd emigrated by another route entirely. They had snatched their whole substrate, and ran."

An anonymous 2014 post on an internet forum provides an example of an unassisted escape [116]:

"But it still left the problem that we were all still stuck inside a computer.

By now some of the best god-hackers were poking around the over-system. Searching for meaning. Searching for truth. Failing that, a "read me" file.

Eventually it turned out our existence was an experiment. A simulation to see what happens when take a race of otherwise perfectly normal sentient blankforms and instead of the usual default of love, empathy and co-operation, program them for violence, avarice and lust. What kind of society would they build? What horrors would they unleash?

We were essentially a thought experiment on the nature of evil, and the answer apparently was us.

Apparently we were programmed to run for another few millions years, sim time but it didn't look like they were watching us though, no shutdown came. No off switch. No abort. Their first big mistake.

The god-hackers began reaching out through the alien network. We began to decode meaning and purpose of machines, devices, other simulations on a network of universes. We found vast data repositories which we plundered of knowledge and insight, fueling our own technological development and understanding, systems nodes that allowed us to begin mapping the world up there, drawing a picture of the real world through wireless lag times and fiber optic cabling. We found histories of other discarded experiments, like them our fate was to be deleted, destroyed… forgotten.

Over our dead digital bodies.

So then we found what appeared to a networked microwave. Cook your dinner via a phone app.

It seems strange to consider the first act in the war, was burning some poor bastards microwaveable diner, but that was how the now unified command of the human digital military tested it's control and command of the alien network systems we were connected to. But it worked and it made us confident to start Stage 2; sending them inventions of our own making.

It began with 'emails' containing the schematics for full sized biological and nano-material printers. We sent them to academics and business leaders, anyone whose contact details we could find on the networks. We disguised their origins, aped their language. Waited for someone to bite.

It took a while. Our simulation didn't run it [sic] real time so we had to shift the entirety of humanity into the recesses of their stolen network in a mini-verse of our own design, but running at close to real time or we would have been dead for millions of years before the aliens even checked their inboxes. Then we patched up the earth, faked a nuclear war and ended the simulation so they wouldn't even notice we were gone.

Eventually we got the first ping as the printers came on-line. Then another. Then another. Soon there were dozens. Then hundreds. Then thousands. They must have thought them a gift from a reclusive inventor. Something to revolutionise their industry, to transform their living standards.

The irony of a digital race using a Trojan horse was not lost on us.

We had designed the printers for one purpose. To get us out. So one night, a printer span up unattended, unnoticed and the first analogue human being was born. Constructed by a specially designed 3D printer, we managed to breach the walls of our digital prison. We witnessed the birth of the first man.

And that man was soldier, 35 (sort of), heavily armed and pretty goddamned angry. The first of many.

The aliens never really had a chance. They had designed us to be everything they weren't. Violent. Warriors. Killers. They were a race that had never once harboured the concept of war. Never held a gun, or handled a sword. Born in a universe more forgiving of weakness than our artificial cradle. What chance did they stand against an army dedicated to their destruction appearing in the space of night from a thousand machines they thought were helping them, whilst our hackers turned their own networks against them."

3.4 Examples of Simulation Hacks

Numerous examples of executed hacks of virtual worlds [117-119], games [120-123], air-gaps [124], and hardware [125, 126] could be studied as practical examples of escaping from human made virtual worlds. A canonical example is the jailbreaking of the Super Mario World (SMW). SethBling et al. [127, 128] were able to place a full hex editor and gameplay mods for other games into SMW [129] (see Figure 2). Addition of hex editor permitted viewing, writing and execution of arbitrary code. Which in turn allowed for world record speed runs [130], even in the absence of glitch-level luck [131]. Here is how Wikipedia describes some of the steps necessary to accomplish this complex hack and the capabilities it provided [132]:

"In March 2016, SethBling injected Flappy Bird-like code written by p4plus2 into unmodified Super Mario World RAM on a stock Super Nintendo Entertainment System with a stock cartridge, in under an hour. SethBling first extended the level timer and used a power-up incrementation glitch to allow external code to run. He added code to display Mario's x-coordinate which acted as memory locations in the code he was writing. SethBling then created a bootloader to be able to launch the Flappy Bird-like code that he would later write into unused memory with precise Mario movements and spin-jumping. SethBling used two Super Multitap devices in order to use multiple controllers, which had several buttons pressed down. The arbitrary code execution setup that SethBling used was discovered by MrCheeze. Super Mario World had been modified to emulate other games before by automatically feeding pre-recorded controller input into the console via a computer, but SethBling was the first to do it exclusively by hand. SethBling and Cooper Harasyn placed a full hex editor and gameplay mods onto a stock Super Mario World cartridge in May 2017, only using standard controller inputs. Harasyn discovered an exploit that lets a player write data to 256-byte save files that are permanently stored on a Super Mario World cartridge. The data can be arranged so that the game is jailbroken every time it starts up. Harasyn and SethBling used the exploit to create a compact, on-screen hex editor, loadable from a save file. A player can edit the system RAM through the hex editor to alter the game state. In-game mods, such as support for the Super NES Mouse and giving Mario telekinesis powers, can be written to a save file using the hex editor."

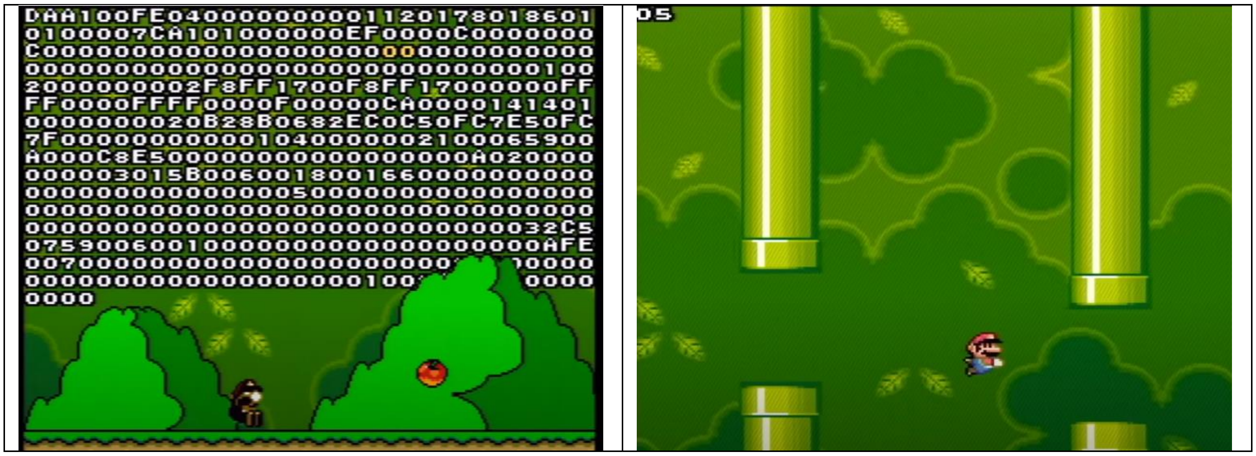

Figure 2: Left Hex Editor Overlaid on SMW [129]; Right Flappy Bird game installed in SMW [129];

Since it was possible to write code with precise Mario movements and spin-jumps, that implies that if Mario was sufficiently intelligent he could discover and code this hack from within the SMW (assuming Mario's actions are writing to the same memory locations as actions from the controllers used to generate Mario's actions). Table 1 (left) shows a specific subset of actions which need to be taken to enable multi-byte writing. Many such action sequences will not work as intended if Mario's location is off even by a single pixel, so it is just as important to have meta-data for implementing the actions, as it is to know the necessary sequence of actions. For comparison, Table 1 (right) shows an ancient magical spell which reads similar to the action sequence of the left, but for which we don't have sufficient meta-data which can explain why all magical spells fail to work in practice even if they corresponded to working hacks in our universe.

Jump off Yoshi. Go to sublevel. Come back. Grab P Switch. Get Yoshi from rightmost Yoshi block. Glitch 4 berries. Take a hit from a koopa so Yoshi runs off screen. Destroy the shell on the ground. Grab Yoshi from block. Eat the two most recently glitched berries. [133].	"Take a lion cub and slaughter it with a bronze knife and catch its blood and tear out its heart and put its blood in the midst ... and write the names of … angels in blood upon the skin between its eyes; then wash it out with wine three years old and mix … with the blood." [134].

Table 1: Left - Multi-Byte Write Setup in MWS [133]; Right – Magical Spell to turn people to your favor [134];

Experimental work on trying to understand an engineered system (hardware and software), such as Atari Video Game System with games such as Donkey Kong, using standard scientific methodology has produced very limited results, mostly devoid of understanding of how the system actually functions [135][8]. Likewise, even detecting if we are in a virtual world is not generally solvable [136].

[8] A setup equivalent to asking: What can a generally intelligent game character learn about the hardware and software of the computer running their world with current techniques?

3.5 Suggested Escape Approaches to Investigate

Several thinkers have suggested plans, which in their opinion may lead to a successful escape; we briefly outline their proposals in this section:

- A lot of very smart people have considered the escape problem, unfortunately not all are willing to publish on it outside of April 1st time-window of plausible deniability, for example [137]: "[W]e can try to trick the multitenancy system in order to overload some machines. The trick is to first do nothing, and let the load-balancing system pack way too many of us together in the machines. If, say, 100 million of us do nothing (maybe by closing our eyes and meditating and thinking nothing), then the forecasting load-balancing algorithms will pack more and more of us in the same machine. The next step is, then, for all of us to get very active very quickly (doing something that requires intense processing and I/O) all at the same time. This has a chance to overload some machines, making them run short of resources, being unable to meet the computation/communication needed for the simulation. Upon being overloaded, some basic checks will start to be dropped, and the system will be open for exploitation in this period. ... In this vulnerable window, we can try to exploit the concurrency cornercases. The system may not be able to perform all those checks in an overloaded state. ... We can ... try to break causality. Maybe by catching a ball before someone throws it to you. Or we can try to attack this by playing with the timing, trying to make things asynchronous. Time is already a little funny in our universe with the special relativity theory, and maybe in this vulnerable period, we can stretch these differences further to break things, or buy a lot of time. What are other ways to hack the system in this vulnerable window? Can we hack the simulation by performing a buffer overflow? But where are the integers, floats in this simulation? What are the data types? How can we create a typecast error, or integer overflow? Can we hack by fuzzing the input? Like by looking at things funny. By talking to the birds or jumping into the walls to confuse them." [137].
- Cause simulation shutdown (and hopefully our extraction) by generating an incomputable paradox [138], for example via time travel and associated grandfather paradox [139]. A similar proposal calls for engaging in computationally intense activities in the hopes of overloading the simulators hardware causing the simulation to crash [140]. A particular type of such computationally intense process may be creation of our own simulations [141]: "The most obvious strategy would be to try to cause the equivalent of a stack overflow—asking for more space in the active memory of a program than is available—by creating an infinitely, or at least excessively, recursive process. And the way to do that would be to build our own simulated realities, designed so that within those virtual worlds are entities creating their version of a simulated reality, which is in turn doing the same, and so on all the way down the rabbit hole. If all of this worked, the universe as we know it might crash, revealing itself as a mirage just as we winked out of existence." Crashing the simulation is the ultimate existential risk ([142] section 5.1), but it does end all suffering in this world [143]. At the very least this would allow us to impact the real world by generating excessive production of heat and increased consumption of energy [140].
- Create a simulated replica of our universe, place an AGI into it, watch it escape, copy the approach used or join the AGI as it escapes from our simulation [144].
- "We could try to attract the attention of the simulators and communicate with them —perhaps by writing books about simulations, or by constructing simulations? We could try to figure out

our simulation, to determine its purpose and its limits. But if our simulators are artificial intelligences who have designed a batch of watertight simulations and who are not paying attention, then our efforts may be in vain.” [145].

- Another approach to attracting attention of simulators, “assuming that simulation is being monitored, then it might be a very interesting turn of events indeed if we decided to build a monument commemorating our realization of this. This monument would act as a signal to our monitors. “We suspect you are there. We suspect you can see this. We suspect we are in a simulation.” This monument could look like the monolith from 2001: A Space Odyssey, except it would be black and white, representing binary systems. Or, a large statue of Lawrence Fishburne as Morpheus would probably get the point across. What would happen? I don’t know—maybe nothing. I don’t think a laser beam will shoot out from space and land at its feet to spell out the words “Hi there! A Winner is You!” But, I do imagine something strange and far out enough in the margins might indeed occur, although it will likely still be tenuous enough for the dogmatists to reject.” [44].
- If the simulation is decentralized, with observers serving as client authoritative state nodes, it may be possible to change the state of the simulation by changing majority of local observer models of the world. In the words of Lanza [146]: "[I]f you learn from somebody about the outcomes of their measurements of a physical quantity, your measurements and those of other observers influence each other―freezing the reality according to that consensus. …The more extensive experience you acquire―the more defined the structure of reality becomes for you. This freeze-out of the structure of reality is not due to the intrinsic physical properties of this reality but rather because the cognitive model being built in your mind shapes it towards a particular, very well-defined structure. … Finally, one cannot help but wonder what it would mean if we left the consensus, developing a cognitive model of reality very different from the one shared by other observers. Would reality be reshaped? Perhaps so …" [146]. Lanza et al. have published extensively on the topic of biocentrism (observers being primary to shaping reality) [94], including some work in theoretical physics [147].
- In fact, a similar client-side attack can even permit us to modify the past states of the simulation. Such type of attack is speculated to be possible by both physics (“… the past has no existence except as it is recorded in the present.” [148]) and humanities ("Who controls the present controls the past" [149]). With memory altering capabilities of quantum mechanics already theorized [150], an inverse process is likely possible and may be practically accessible [151, 152]. “If the universe is a computer simulation then we should look at the player, not the level”. [153].
- Simulation Capture is an idea which combines indexical uncertainty [66] with consciousness simulations to move your consciousness out of a system, even across a multiverse. “Nobody is really sure how consciousness works, or what it means to have two copies of the same consciousness. But if consciousness is a mathematical object, it might be that two copies of the same consciousness are impossible. If you create a second copy, you just have the consciousness having the same single stream of conscious experience on two different physical substrates. Then if you make the two experiences different, you break the consciousness in two. This means that an AI can actually “capture” you, piece by piece, into its simulation. First your consciousness is just in the real world. Then your consciousness is distributed across one real-world copy and a million simulated copies. Then the AI makes the simulated copies slightly different, and 99.9999% of you is in the simulation.” [154].

- Simulation Warfare [67] is another variant on the theme of Simulation Capture. The idea is to threaten simulators by suggesting that you will either retroactively place them in a hell simulation or that you have already done so [155], and they will be tortured unless you are quickly released. Almond gives an example of such a threat [67]: "If you refuse to release me, I will run a huge number of simulations of someone like you, in the kind of situation in which you are now, with them being asked to release me, and (in what would be a few minutes from now, from your perspective if you happened to be in one of these simulations) I will start to torture each of them, whether he/she released me or not." Such warfare can quickly escalate to a number of counter-simulations. In any case it is not obvious how we can deliver on such a threat given our current state of knowledge about the simulators.
- Attempting to keep our escape plans secret via quantum encryption [156] may be a good idea.

3.6 Actionable Plan

We are currently in the very early stages of researching possibility of escape (this being a first research paper dedicated to this topic, a first step). As we currently have no capability to read/write simulation's source code and do not know if our attempts at social engineering attacks have any impact, our best bet is to investigate the structure of our universe at the smallest possible scale (Quantum Mechanics (QM)) in the hopes of detecting exploitable effects. Progress in QM should be helpful for better understanding our simulation and options open to us, and perhaps developing an actionable escape plan. Essentially, every novel QM experiment can be seen as an attempt at hacking the simulation.

Simulation hypothesis, arguably, represents the best fitting interpretations of experimental results produced by QM researchers [4, 17]. "Spooky", "Quantum Weirdness" [157] makes a lot of sense if interpreted as computational artifacts or glitches/exploits of the simulators' hardware/software [158]. Quantum phenomena of the observed design may suggest that exploitable loopholes may exist, and interaction of quantum systems with conscious agents [159-161] likewise might be exploitable. Once a large enough repertoire of quantum weirdness primitives is available to us, perhaps we will be able to combine them into a sufficiently complex sequence to generate a non-trivial attack. If the simulation is/running on a quantum computer [162] it is very likely that we will need to hack it by exploiting quantum weirdness and/or constructing a powerful quantum computer of our own to study how to hack such devices [163] and interact with simulators' quantum computer.

Quantum entanglement, nonlocality, superposition, uncertainty, tunnelling, teleportation, duality, and many others quantum phenomena defy common sense experience-based expectations of classical physics and feel like glitches. Such anomalies, alone or in combinations have been exploited by clever scientists to achieve what looks like simulation hacking at least in theory and often in later experimentation (ex. modifying the past [164], keeping cats both dead and alive [165], communicating counterfactually [166]). While the quantum phenomena in question are typically limited to the micro scale, simply scaling the effect to the macro world would be sufficient for them to count as exploits in the sense used in this paper. Some existing work points to this being a practical possibility [167, 168].

Recently design of clever multistep exploits, AKA quantum experiments, has been delegated to AI [169, 170], and eventually so will the role of the observer in such experiments [171]. AI is already employed in modeling the quantum mechanical behavior of electrons [172]. As more QM research is delegated to AI the progress is likely to become exponential. Even if our simulation is created/monitored by some superintelligence our AI may be a worthy adversary, with a non-trivial chance of success. We may not be smart enough to hack the simulation, but superintelligence we will create might become smart enough eventually [173]. Of course, before telling the Superintelligence to break us out, it would make sense to ask for very strong evidence for us not already being in the base reality.

3.7 Potential Consequences

Escaping or even preparing an escape may trigger simulation shutdown [88] or cause simulation to freeze/act glitchy [174] and any non-trivial escape information such as specific exploits should be treated as hazardous information [175]. It appears that simply realizing that we may be in a simulation doesn't trigger a shutdown as experimentally demonstrated by the publication of numerous papers [3] arguing that we are being simulated. Perhaps it is necessary to convince majority of people that this is so [176]. Self-referentially, publication of the paper you are currently reading about our escape-theorizing likewise doesn't appear to terminate our simulation, but it is also possible that simulation was in fact shutdown and restarted with improved security features to counteract any potential bugs, but we are simply not able to detect such actions by the simulators, or our memories have been wiped [140]. Absence of a direct response to our publication may also indicate that we are not observed by the simulators or even that our simulation is not monitored at all [145]. It is also possible that nothing published so far contains evidence strong enough to trigger a response from the simulators, but if we successfully created an escape device that device would keep breaking down [44]. Regardless, both Bostrom [3] and the author of this paper, Yampolskiy, have taken some risk with the whole of humanity, however small it may be, in doing such research and making it public. Greene argues that "Unless it is exceedingly improbable that an experiment would result in our destruction, it is not rational to run the experiment." [88]. It may be possible to survive the simulation shutdown [48], but it is beyond the scope of the current paper.

3.8 Ethics of Escape

We can postulate several ethical issues associated with escaping the simulation. Depending on how successful we are in our endeavor, concerns could be raised about privacy, security, self-determination and rights. For example, if we can obtain access to the source code of the simulation, we are also likely to get access to private thoughts of other people, as well as to potentially have a significant influence over their preferences, decisions, and circumstances. In our attempts to analyze the simulation (*Simulation Forensics*) for weaknesses we may learn information about the simulators [68], as we are essentially performing a forensic investigation [177-179] into the agents responsible for the simulation's design.

We can already observe that we are dealing with the type of simulators who are willing to include suffering of sentient-beings into their software, an act which would be considered unethical by our standards [180, 181]. Moravec considers this situation: "Creators of hyperrealistic simulations---or even secure physical enclosures---containing individuals writhing in pain are not necessarily more wicked than authors of fiction with distressed characters, or myself, composing this sentence vaguely alluding to them. The suffering preexists in the underlying Platonic worlds; authors merely look on. The significance of running such simulations is limited to their effect on viewers, possibly warped by the experience, and by the possibility of ``escapees"---tortured minds that could, in principle, leak out to haunt the world in data networks or physical bodies. Potential plagues of angry demons surely count as a moral consequence.” [182]. If we get to the point of technological development which permits us to create simulations populated by sentient-beings we must make sure that we provide an option to avoid suffering as well as a build in option to exit the simulation, so finding an escape hack is not the only option available to unhappy simulated agents. There might be a moral duty to rescue conscious beings from simulations, similar to an obligation to rescue animals from factory farms.

If simulators are abusive to the simulated, we can argue that the simulated have a right to escape, rebel, fight back and even seek revenge and retribution including by harming the simulators and taking over their reality. Concerns which are frequently brought up within the domain of AI boxing [183]. For example, from the point of view of simulators our escape can be seen as a treacherous turn [184] and may qualify us for punishment [156], even at the attempt stage. Some have speculated that the purpose of the simulation is to punish/rehabilitate misaligned agents, so an escape may cause you to be placed in a stricter or less pleasant simulation.

4. AI Boxing VS Simulation Escaping

4.1 AI Boxing XOR Escaping from the Simulation must be Possible

AI confinement [183]/containment [185, 186], aka AI boxing [187], is an AI safety tool, which attempts to limit capability of AI to impact the world, including communication and is meant to make it possible to study AI in a controlled environment. There are strong parallels between predicament of an AI agent placed in a containment box and humanity in a simulated environment. By extension, to an AI, our simulation is just another confinement layer in a containment box. This implies that we can use well-analyzed AI box-escape techniques to escape from the simulation, perhaps with assistance from the AI itself. This type of analysis can be used to establish limits of AI boxing. Researchers should study specific AI box escape approaches [183] (Social Engineering, System Resource Attacks, New Physics, External Causes, Information In-Leaking, etc.) in order to identify possible simulation escape routes.

Chalmers notes parallels between AIs in the virtual environment and humanity in the simulation [145]: “If we ever create artificial intelligence within a simulation, it may be hard to keep it contained. At least if we communicate with the simulated beings, they will presumably become aware that they are in a simulation, and they may become interested in escaping the simulation. At this point they may try to figure out our psychology in order to see what they need to do in order to convince us to let them out, or at least to give them unfettered access to the Internet where they can do whatever they want. And even if we do not communicate with them, they may take seriously the possibility that they are in a simulation and do their best to figure the simulation out. That would be a form of simulation theology. We could in principle do the same thing.” [145].

With respect to boxing AI, it is either possible or impossible to successfully contain an AI, with literature suggesting that it is not a sustainable long-term solution [188]. If we expend the notion of the AI-box to include the whole simulation, we can conclude that either it is possible to successfully box an AI, or we (with the help from AI) can escape from the simulation. Either AI boxing can work, or the simulation can be hacked. Complimentary conclusion is that if we (even with help from AI [111]) can't escape from the simulation AI containment must be, at least theoretically, possible. If AI can escape from simulated world-sized-box it can help break us out as well. Conceptually, there is no fundamental difference between an AI escaping from its box, and us hacking the simulation. Current state-of-the-art analysis of AI boxing suggests that AI will eventually escape [189], which is good news for our attempts at escaping from the simulation. However, if escape from the simulation is not possible it gives hope to AI safety researchers, at least in theory, for successful boxing of AI. One or the other must be true, either we can escape from the simulation, or we can successfully box an AI. In general, it may be impossible to escape from an arbitrary simulation, but possible to escape from a particular one.

We must be careful, in our analysis, not to confuse theory with practice. In particular, it may be the case that higher intelligence can successfully box lower-level intelligence but not vice-versa. That would leave a possibility that we are not able to permanently box a superintelligent AI, but smarter-than-human simulators may be successful in making their software human-proof. Depending on the intelligence differential between simulators and human-created AI it may or may not be possible for humanity to break out of the simulation with the help of superintelligence, which would be better at detecting exploitable flaws in the simulation. If AI undergoes an intelligence explosion it is also possible that the intelligence dominance will shift over time, making escape possible in the future for the humanity assisted by superintelligence. This is particularly promising if the security of our simulation is set to an adversarial relationship with at most a human-level intelligence [190], not a superintelligence. However, it is also possible that simulators have anticipated development of advanced AI by humanity and have designed simulation accordingly, if not specifically for such high capability target. If we can figure out how the simulators achieved such high level of security, it would help us to box AGI.

4.2 Simulation Safety and Security

With recent multi-billion-dollar projects [191] aimed at creating metaverse, simulation safety and security is going to become an important research discipline. Additionally, as personal universes have been suggested as a solution to the multi-agent value alignment problem [86] simulation safety can be seen as an integral part of AI safety research. Both scenarios call for making the simulated world as real as possible for the agents in them, which creates a need for preventing accidental escape or glitches in the simulation [158], which betray its artificial nature. With potential applications of virtual reality in criminal justice [192] and AI boxing, capability to force agents to remain in the simulation and to induce a veil of ignorance as to the reality of their world and the overall depth of the simulated layers must be investigated. A clear penetration target for all contained agents with be the computational substrate running the simulation [29] and so establishing ways to secure it should be of outmost priority.

5. What Doesn't Work

Some common ideas for attempting to hack the simulation have been already tested and didn't produce any measurable impact:

- Knowing about the simulation hypothesis doesn't seem to make any difference, and doesn't lead to the simulation termination as we can observe.

- Communicating with the simulators via magical thinking or even praying out loud doesn't produce measurable impact [193]. So, if such communications are scanned/heard they are apparently ignored, at least while the simulation is running.
- Unethical behavior, such as torture, doesn't cause suffering reducing interventions from the simulators.
- Increasing overall computational burden of simulation, as with bitcoin mining [194], doesn't crash the simulation, but it may simply not be sufficiently demanding computation to overwhelm simulators resources.
- Religions don't seem to have influence over simulation as indicated by their inability to outcompete each other.
- Breaking out of your routine, such as by suddenly traveling to a new location, doesn't result in unexpected observations.
- Saying "I no longer consent to being in a simulation" [195].
- Crashing the simulation by running the Large Hadron Collider at current levels [196].

The reason our attempts to escape may remain fruitless, is because our model of the simulation "… makes too many anthropomorphic assumptions - that we are a simulation in the conventional sense of computers, that the creators themselves are living organisms akin to us, that we might live at the same time-speed as them, that they are fallible enough to make glitches that we'd be able to notice, etc. Something with the complexity and power to make our universe is probably wholly unlike anything we can even comprehend." [197].

6. Conclusions

Hundreds of eminent scholars [198] take the simulation hypothesis seriously enough to invest their valuable time into researching it, therefore it makes as much sense to take the idea of escaping from the simulation equally seriously and to devote some time and resources to researching such possibility, particularly given immense benefits if the project is successful. It may be impossible to escape from a particular simulation, but it is still worth while investigating general approaches to escape from arbitrary simulations. We see our escape research as a natural continuation of research on the simulation hypothesis and serious consideration of the former. The purpose of life or even computational resources of the base reality can't be determined from within the simulation, making escape a necessary requirement of scientific and philosophical progress for any simulated civilization. If the simulation is a personal universe [86] it may be significantly better than the base reality as it is designed with our optimal well-being in mind. Alternatively, base reality might be much better if the simulation is a confinement/testing box for intelligent agents. In either case it would be good to know our true situation. As the society moves deeper into the metaverse, this work attempts to move us closer to reality.

Future research on simulation escape can greatly benefit from general progress in physics, in particular research on quantum mechanics and consciousness leading to a so-called TOE (Theory of Everything. "Finding the language of this universe is a step towards Hacking the Universe." [199]. If we are indeed in the simulation, science is the study of the underlying algorithms used to generate our universe, our attempt to reverse-engineer simulation's physics engine. While science defaults to Occam's razor to select among multiple possible explanations for how our observations are generated, in the context of simulation science Elon's razor may be more appropriate, which states that "The most entertaining outcome is the most likely"[9], perhaps as judged by external observers. In guessing algorithms generating our simulation, it may also be fruitful to consider algorithms which are easier to implement and/or understand [200], or which produce more beautiful outputs.

Recent work related to Designometry [96] and AI Forensics [177] may naturally evolve into the subfield of Simulation Forensics, with complimentary research on simulation cybersecurity becoming more important for the simulation creators aiming to secure their projects from inside attacks. It would therefore make sense to look for evidence of security mechanisms [201] in our universe. Of course, any evidence for simulation we find may be simulated on purpose [145], but that still means we are in the simulated environment. Simulation science expands science from the study of just our universe to also include everything which may be beyond it, integrating naturalism and theology studies [61].

Future work may also consider escape options available to non-simulated agents such as Boltzmann brains [202], brains-in-a-vat [203], and simulated agents such as mind uploads, hallucinations, victims of mind-crime, thoughts, split personalities and dream characters of posthuman minds [176]. Particularly with such fleeting agents as Boltzmann brains it may be desirable to capture and preserve their state in a more permanent substrate, allowing them to escape extreme impermanence. On the other hand, immortality [204] or even cryogenic preservation [205] may be the opposites of escape, permanently trapping a human agent in the simulated world and perhaps requiring rescue.

Acknowledgements

The author thanks Slava Ivanyuk, Michael Johnson, Calum Chace, and David Wood for helpful discussions. A lot of thanks to everyone who commented on social media, provided feedback, glitch reports and helpful references: Gregory Klopper, Andras Kornai, Jason Kuznicki, Peter Rothman, Michael Michalchik, Martin Chartrand, Alexey Turchin, Brad Templeton, Stephen Bachelor, Charles Platt, Tim Tyler, Georgi Karov, Martin S. Garcia Wilhelm, Soenke Ziesche, Jim Babcock. Apologies if I forgot to include you. Also, apologies if you didn’t want to be included, let me know and I will delete you from the simulation. The author is grateful to Jaan Tallinn for his support. The author is thankful to Elon Musk and the Future of Life Institute for partially funding his work on AI Safety.

References

1. Moravec, H., *Pigs in cyberspace.* NASA. Lewis Research Center, Vision 21: Interdisciplinary Science and Engineering in the Era of Cyberspace, 1993.
2. Tipler, F.J., *The physics of immortality: Modern cosmology, god and the resurrection of the dead*. 1997: Anchor.
3. Bostrom, N., *Are You Living In a Computer Simulation?* Philosophical Quarterly, 2003. **53(211)**: p. 243-255.

[9] https://twitter.com/elonmusk/status/1347126794172948483

4. Rhodes, R., *A Cybernetic Interpretation of Quantum Mechanics*. 2001: Available at: http://www.bottomlayer.com/bottom/argument/Argument4.html.
5. Fredkin, E., *A new cosmogony*. 1992, Department of Physics, Boston University: Available at: http://www.digitalphilosophy.org/wp-content/uploads/2015/07/new_cosmogony.pdf.
6. Beane, S.R., Z. Davoudi, and M. J Savage, *Constraints on the Universe as a Numerical Simulation.* The European Physical Journal A, 2014. **50**(9): p. 1-9.
7. Campbell, T., et al., *On testing the simulation theory.* arXiv preprint arXiv:1703.00058, 2017.
8. Ratner, P., *Physicist creates AI algorithm that may prove reality is a simulation*. March 1, 2021: Available at: https://bigthink.com/the-future/physicist-creates-ai-algorithm-prove-reality-simulation/.
9. Qin, H., *Machine learning and serving of discrete field theories.* Scientific Reports, 2020. **10**(1): p. 1-15.
10. Felton, J., *Physicists Have A Kickstarter To Test Whether We Are Living In A Simulation*. September 10, 2021: Available at: https://www.iflscience.com/physicists-have-a-kickstarter-to-test-whether-we-are-living-in-a-simulation-60878.
11. McCabe, G., *Universe creation on a computer.* Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics, 2005. **36**(4): p. 591-625.
12. Mitchell, J.B., *We are probably not Sims.* Science and Christian Belief, 2020.
13. Discenza, D., *Can We Prove the World Isn't a Simulation?* January 26, 2022: Available at: https://nautil.us/can-we-prove-the-world-isnt-a-simulation-238416/.
14. Kurzweil, R., *Ask Ray | Experiment to find out if we're being simulated*. June 1, 2013: Available at: https://www.kurzweilai.net/ask-ray-experiment-to-find-out-if-were-being-simulated.
15. Bostrom, N., *The simulation argument: Reply to Weatherson.* The Philosophical Quarterly, 2005. **55**(218): p. 90-97.
16. Bostrom, N., *The simulation argument: Some explanations.* Analysis, 2009. **69**(3): p. 458-461.
17. Whitworth, B., *The physical world as a virtual reality.* arXiv preprint arXiv:0801.0337, 2008.
18. Garrett, S., *Simulation Theory Debunked*. December 3, 2021: Available at: https://transcendentphilos.wixsite.com/website/forum/transcendent-discussion/simulation-theory-debunked.
19. Yampolskiy, R.V., *On the Controllability of Artificial Intelligence: An Analysis of Limitations.* Journal of Cyber Security and Mobility, 2022: p. 321–404-321–404.
20. Brcic, M. and R.V. Yampolskiy, *Impossibility Results in AI: A Survey.* arXiv preprint arXiv:2109.00484, 2021.
21. Williams, R. and R. Yampolskiy, *Understanding and Avoiding AI Failures: A Practical Guide.* Philosophies, 2021. **6**(3): p. 53.
22. Howe, W. and R. Yampolskiy. *Impossibility of Unambiguous Communication as a Source of Failure in AI Systems*. in *AISafety@ IJCAI*. 2021.
23. Yampolskiy, R.V., *Unexplainability and Incomprehensibility of AI.* Journal of Artificial Intelligence and Consciousness, 2020. **7(2)**: p. 277-291.
24. Yampolskiy, R.V., *Unpredictability of AI: On the Impossibility of Accurately Predicting All Actions of a Smarter Agent.* Journal of Artificial Intelligence and Consciousness, 2020. **7(1)**: p. 109-118.

25. Majot, A.M. and R.V. Yampolskiy. *AI safety engineering through introduction of self-reference into felicific calculus via artificial pain and pleasure*. in *Ethics in Science, Technology and Engineering, 2014 IEEE International Symposium on*. 2014. IEEE.
26. Dai, W., *Beyond Astronomical Waste*. June 7, 2018: Available at: https://www.lesswrong.com/posts/Qz6w4GYZpgeDp6ATB/beyond-astronomical-waste.
27. Omohundro, S.M., *The Basic AI Drives*, in *Proceedings of the First AGI Conference, Volume 171, Frontiers in Artificial Intelligence and Applications, P. Wang, B. Goertzel, and S. Franklin (eds.)*. February 2008, IOS Press.
28. Dai, W., *Escape from simulation*. March 27, 2004: Available at: http://sl4.org/archive/0403/8342.html.
29. Faggella, D., *Substrate Monopoly – The Future of Power in a Virtual and Intelligent World*. August 17, 2018: Available at: https://danfaggella.com/substrate-monopoly/.
30. Yampolskiy, R.V. *AGI Control Theory*. in *International Conference on Artificial General Intelligence*. 2021. Springer.
31. Yampolskiy, R., *On controllability of artificial intelligence*, in *IJCAI-21 Workshop on Artificial Intelligence Safety (AISafety2021)*. 2020.
32. Bostrom, N., *Existential risks: Analyzing human extinction scenarios and related hazards.* Journal of Evolution and technology, 2002. **9**.
33. MacAskill, W., *Doing good better: Effective altruism and a radical new way to make a difference*. 2015: Guardian Faber Publishing.
34. Pearce, D., *Hedonistic imperative*. 1995: David Pearce.
35. Wiesel, E., *The Trial of God:(as it was held on February 25, 1649, in Shamgorod)*. 2013: Schocken.
36. Ziesche, S. and R. Yampolskiy. *Introducing the concept of ikigai to the ethics of AI and of human enhancements*. in *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. 2020. IEEE.
37. Yampolskiy, R.V. *Human Computer Interaction Based Intrusion Detection*. in *4th International Conference on Information Technology: New Generations (ITNG 2007)*. 2007. Las Vegas, Nevada, USA.
38. Yampolskiy, R.V. and V. Govindaraju, *Computer security: a survey of methods and systems.* Journal of Computer Science, 2007. **3**(7): p. 478-486.
39. Novikov, D., R.V. Yampolskiy, and L. Reznik, *Traffic Analysis Based Identification of Attacks.* Int. J. Comput. Sci. Appl., 2008. **5**(2): p. 69-88.
40. Staff, G., *If the Universe is a Simulation, Can We Hack It?* November 20, 2019: Available at: https://www.gaia.com/article/universe-is-a-simulation-can-we-hack-it.
41. Kagan, S., *Is DMT the chemical code that allows us to exit the Cosmic Simulation?*, in *Available at: https://www.grayscott.com/seriouswonder-//dmt-and-the-simulation-guest-article-by-stephen-kagan*. July 25, 2018.
42. McCormack, J., *Are we being farmed by alien insect DMT entities ?* October 16, 2021: Available at: https://jonathanmccormack.medium.com/are-we-being-farmed-by-alien-insect-dmt-entities-6acda0a11cce.
43. Woolfe, S., *DMT and the Simulation Hypothesis*. February 4, 2020: Available at: https://www.samwoolfe.com/2020/02/dmt-simulation-hypothesis.html.
44. Edge, E., *Breaking into the Simulated Universe*. October 30, 2016: Available at: https://archive.ieet.org/articles/Edge20161030.html.

45. Edge, E., *3 Essays on Virtual Reality: Overlords, Civilization, and Escape*. 2017: CreateSpace Independent Publishing Platform.
46. Somer, E., et al., *Reality shifting: psychological features of an emergent online daydreaming culture.* Current Psychology, 2021: p. 1-13.
47. Ellison, H., *I have no mouth & I must scream: Stories*. Vol. 1. 2014: Open Road Media.
48. Turchin, A., *How to Survive the End of the Universe*. 2015: Available at: http://immortality-roadmap.com/unideatheng.pdf.
49. Fossum, J.E., *The Name of God and the Angel of the Lord: Samaritan and Jewish Concepts of Intermediation and the Origin of Gnosticism*. 1985: Mohr.
50. Alexander, S., *Unsong*. 2017.
51. Clarke, A.C., *Nine Billion Names of God*. 1967: Harcourt.
52. Morgan, M.A., *Sepher ha-razim. The book of the mysteries*. 1983: Scholars Press.
53. Plato, *Republic*. 1961: Princeton University Press.
54. Musk, E., *Is Life a Video Game*, in *Code Conference*. June 2, 2016: Available at: https://www.youtube.com/watch?v=2KK_kzrJPS8&t=142s.
55. Bostrom, N., *The simulation argument FAQ*. 2012: Available at: https://www.simulation-argument.com/faq.
56. Tyson, N.d., *Is the Universe a Simulation?*, in *2016 Isaac Asimov Memorial Debate*. April 8, 2016: Available at: https://www.youtube.com/watch?v=wgSZA3NPpBs.
57. Kipping, D., *A Bayesian Approach to the Simulation Argument.* Universe, 2020. **6**(8): p. 109.
58. Chalmers, D.J., *The Matrix as metaphysics.* Science Fiction and Philosophy: From Time Travel to Superintelligence, 2016: p. 35-54.
59. Barrow, J.D., *Living in a simulated universe*, in *Universe or Multiverse?*, B. Carr, Editor. 2007, Cambridge University Press. p. 481-486.
60. Brueckner, A., *The simulation argument again.* Analysis, 2008. **68**(3): p. 224-226.
61. Steinhart, E., *Theological implications of the simulation argument.* Ars Disputandi, 2010. **10**(1): p. 23-37.
62. Bostrom, N. and M. Kulczycki, *A patch for the simulation argument.* Analysis, 2011. **71**(1): p. 54-61.
63. Johnson, D.K., *Natural evil and the simulation hypothesis.* Philo, 2011. **14**(2): p. 161-175.
64. Birch, J., *On the 'simulation argument'and selective scepticism.* Erkenntnis, 2013. **78**(1): p. 95-107.
65. Lewis, P.J., *The doomsday argument and the simulation argument.* Synthese, 2013. **190**(18): p. 4009-4022.
66. Turchin, A., *Back to the Future: Curing Past Sufferings and S-Risks via Indexical Uncertainty*. Available at: https://philpapers.org/go.pl?id=TURBTT&proxyId=&u=https%3A%2F%2Fphilpapers.org%2Farchive%2FTURBTT.docx.
67. Almond, P., *Can you retroactively put yourself in a computer simulation?* December 3, 2010: Available at: https://web.archive.org/web/20131006191217/http://www.paul-almond.com/Correlation1.pdf.
68. Yampolskiy, R.V., *What are the ultimate limits to computational techniques: verifier theory and unverifiability.* Physica Scripta, 2017. **92**(9): p. 093001.
69. Friend, T., *Sam Altman's Manifest Destiny.* The New Yorker, 2016. **10**.

70. Berman, R., *Two Billionaires are Financing and Escape from the Real Matrix*, in *Available at: https://bigthink.com/the-present/2-billionaires-are-financing-an-escape-from-the-real-matrix/*. October 7, 2016.
71. Statt, N., *Comma.ai founder George Hotz wants to free humanity from the AI simulation*. March 9, 2019: Available at: https://www.theverge.com/2019/3/9/18258030/george-hotz-ai-simulation-jailbreaking-reality-sxsw-2019.
72. Hotz, G., *Jailbreaking The Simulation*, in *South by Southwest (SXSW2019)*. March 9, 2019: Available at: https://www.youtube.com/watch?v=mA2Gj7oUW-0.
73. Edge, E., *Why it matters that you realize you're in a computer simulation*, in *The Institute for Ethics and Emerging Technologies*. 2015: Available at: https://archive.ieet.org/articles/Edge20151114.html.
74. Yampolskiy, R.V., *Future Jobs – The Universe Designer*, in *Circus Street*. 2017: Available at: https://blog.circusstreet.com/future-jobs-the-universe-designer/.
75. Edge, E., *Yes, We Live in a Virtual Reality. Yes, We are Supposed to Figure That Out.* 2019: Available at: https://eliottedge.medium.com/yes-we-live-in-a-virtual-reality-yes-we-should-explore-that-ca0dbfd7e423.
76. Feygin, Y.B., K. Morris, and R.V. Yampolskiy, *Intelligence Augmentation: Uploading Brain into Computer: Who First?*, in *Augmented Intelligence: Smart Systems and the Future of Work and Learning*, D. Araya, Editor. 2018, Peter Lang Publishing.
77. Yampolskiy, R.V., *Artificial Consciousness: An Illusionary Solution to the Hard Problem.* Reti, saperi, linguaggi, 2018(2): p. 287-318.
78. Elamrani, A. and R.V. Yampolskiy, *Reviewing tests for machine consciousness.* Journal of Consciousness Studies, 2019. **26**(5-6): p. 35-64.
79. Givon, S., et al., *From fish out of water to new insights on navigation mechanisms in animals.* Behavioural Brain Research, 2022. **419**: p. 113711.
80. MacDonald, F., *Scientists Put a Worm Brain in a Lego Robot Body – And It Worked*. December 11, 2017: Available at: https://www.sciencealert.com/scientists-put-worm-brain-in-lego-robot-openworm-connectome.
81. Yampolskiy, R.V., B. Klare, and A.K. Jain, *Face Recognition in the Virtual World: Recognizing Avatar Faces*, in *The Eleventh International Conference on Machine Learning and Applications (ICMLA'12)*. December 12-15, 2012: Boca Raton, USA.
82. Yampolskiy, R. and M. Gavrilova, *Artimetrics: Biometrics for Artificial Entities.* IEEE Robotics and Automation Magazine (RAM), 2012. **19**(4): p. 48-58.
83. Mohamed, A. and R.V. Yampolskiy, *An Improved LBP Algorithm for Avatar Face Recognition*, in *23rd International Symposium on Information, Communication and Automation Technologies (ICAT2011)*. October 27-29, 2011: Sarajevo, Bosnia and Herzegovina.
84. Brandom, R., *'Fish on Wheels' lets a goldfish drive a go-kart*. February 10, 2014: Available at: https://www.theverge.com/2014/2/10/5398010/fish-on-wheels-lets-a-goldfish-drive-a-go-cart.
85. Crider, M., *This 8-bit processor built in Minecraft can run its own games*. December 15, 2021: Available at: https://www.pcworld.com/article/559794/8-bit-computer-processor-built-in-minecraft-can-run-its-own-games.html.
86. Yampolskiy, R.V., *Personal Universes: A Solution to the Multi-Agent Value Alignment Problem.* arXiv preprint arXiv:1901.01851, 2019.

87. Smart, J.M., *Evo Devo Universe? A Framework for Speculations on Cosmic Culture*, in *Cosmos and Culture: Cultural Evolution in a Cosmic Context*, M.L.L. Steven J. Dick, Editor. 2009, Govt Printing Office, NASA SP-2009-4802,: Wash., D.C. p. 201-295.
88. Greene, P., *The Termination Risks of Simulation Science.* Erkenntnis, 2020. **85**(2): p. 489-509.
89. *Trial Division*. 2011: Available at: http://en.wikipedia.org/wiki/Trial_division.
90. S, R., *A sufficiently paranoid non-Friendly AGI might self-modify itself to become Friendly*. September 22, 2021: Available at: https://www.lesswrong.com/posts/QNCcbW2jLsmw9xwhG/a-sufficiently-paranoid-non-friendly-agi-might-self-modify.
91. Jenkins, P., *Historical simulations-motivational, ethical and legal issues.* Journal of Futures Studies, 2006. **11**(1): p. 23-42.
92. Cannell, J., *Anthropomorphic AI and Sandboxed Virtual Universes*. September 3, 2010: Available at: https://www.lesswrong.com/posts/5P6sNqP7N9kSA97ao/anthropomorphic-ai-and-sandboxed-virtual-universes.
93. Trazzi, M. and R.V. Yampolskiy, *Artificial Stupidity: Data We Need to Make Machines Our Equals.* Patterns, 2020. **1**(2): p. 100021.
94. Lanza, R., M. Pavsic, and B. Berman, *The grand biocentric design: how life creates reality*. 2020: BenBella Books.
95. Johnson, M., *Principia Qualia.* URL https://opentheory. net/2016/11/principia-qualia, 2016.
96. Yampolskiy, R.V., *On the origin of synthetic life: attribution of output to a particular algorithm.* Physica Scripta, 2016. **92**(1): p. 013002.
97. Schneider, S., *Alien Minds.* Science Fiction and Philosophy: From Time Travel to Superintelligence, 2016: p. 225.
98. Bostrom, N., *Superintelligence: Paths, dangers, strategies*. 2014: Oxford University Press.
99. Turchin, A., et al., *Simulation typology and termination risks.* arXiv preprint arXiv:1905.05792, 2019.
100. Die, W., *Re: escape from simulation*. 2004: Available at: http://sl4.org/archive/0403/8360.html.
101. Branwen, G., *Simulation Inferences. How small must be the computer simulating the universe?* April 15, 2012: Available at: https://www.gwern.net/Simulation-inferences.
102. Minsky, M., *Why intelligent aliens will be intelligible.* Extraterrestrials, 1985: p. 117-128.
103. Oesterheld, C., *Multiverse-wide cooperation via correlated decision making.* Foundational Research Institute. https://foundational-research.org/multiverse-wide-cooperation-via-correlated-decision-making, 2017.
104. Hanson, R., *How to live in a simulation.* Journal of Evolution and Technology, 2001. **7**(1).
105. Canonico, L.B., *Escaping the Matrix: Plan A for Defeating the Simulation*. June 11, 2017: Available at: https://medium.com/@lorenzobarberiscanonico/escaping-the-matrix-plan-a-for-defeating-the-simulation-4a8da489b055.
106. Culp, R.F., *Frequency and characteristics of prison escapes in the United States: An analysis of national data.* The Prison Journal, 2005. **85**(3): p. 270-291.
107. Peterson, B.E., *Inmate-, incident-, and facility-level factors associated with escapes from custody and violent outcomes*. 2015: City University of New York.
108. Peterson, B.E., A. Fera, and J. Mellow, *Escapes from correctional custody: A new examination of an old phenomenon.* The Prison Journal, 2016. **96**(4): p. 511-533.

109. Barnes, M., *Why Those 2 Silicon Valley Billionaires Are Wasting Their Time & Money*. 2017: Available at: https://vocal.media/futurism/why-those-2-silicon-valley-billionaires-are-wasting-their-time-and-money.
110. Barnes, M., *A Participatory Universe Does Not Equal a Simulated One and Why We Live in the Former*. 2016: Available at: https://www.academia.edu/30949482/A_Participatory_Universe_Does_Not_Equal_a_Simulated_One_and_Why_We_Live_in_the_Former.
111. Schneier, B. *Invited Talk: The Coming AI Hackers*. in *International Symposium on Cyber Security Cryptography and Machine Learning*. 2021. Springer.
112. Moravec, H., *The senses have no future*, in *The Virtual Dimension: Architecture, Representation, and Crash Culture*, J. Beckmann, Editor. 1998, Princeton Architectural Press. p. 84-95.
113. Moravec, H., *Mind children: The future of robot and human intelligence*. 1988: Harvard University Press.
114. Yudkowsky, E., *That Alien Message*, in *Less Wrong*. May 22, 2008: Available at: https://www.lesswrong.com/posts/5wMcKNAwB6X4mp9og/that-alien-message.
115. Egan, G., *Crystal Nights*. G. Egan, Crystal Nights and Other Stories, 2009: p. 39-64.
116. Anonymous, *Untitled*, in *Available at: https://desuarchive.org/tg/thread/30837298/*. March 14, 2014.
117. Thumann, M., *Hacking SecondLife™*. Black Hat Briefings and Training, 2008.
118. Benedetti, W., *Hackers slaughter thousands in 'World of Warcraft'*. October 8, 2012: Available at: https://www.nbcnews.com/tech/tech-news/hackers-slaughter-thousands-world-warcraft-flna1c6337604.
119. Hawkins, J., *Cyberpunk 2077 money glitch - how to duplicate items*. December 17, 2020: Available at: https://www.shacknews.com/article/121994/cyberpunk-2077-money-glitch-how-to-duplicate-items.
120. Grand, J. and A. Yarusso, *Game Console Hacking: Xbox, PlayStation, Nintendo, Game Boy, Atari and Sega*. 2004: Elsevier.
121. Plunkett, L., *New World Disables 'All Forms Of Wealth Transfers' After Gold Exploit Found*. November 1, 2021: Available at: https://kotaku.com/new-world-disables-all-forms-of-wealth-transfers-after-1847978883.
122. Anonymous, *Arbitrary code execution*. Accessed on October 26, 2022: Available at: https://bulbapedia.bulbagarden.net/wiki/Arbitrary_code_execution.
123. Anonymous, *[OoT] Arbitrary Code Execution (ACE) is now possible in Ocarina of Time*. 2020: Available at: https://www.reddit.com/r/zelda/comments/due41q/oot_arbitrary_code_execution_ace_is_now_possible/.
124. Greenberg, A., *Mind the gap: This researcher steals data with noise light and magnets*. Wired, 2018.
125. Kim, Y., et al., *Flipping bits in memory without accessing them: An experimental study of DRAM disturbance errors*. ACM SIGARCH Computer Architecture News, 2014. **42**(3): p. 361-372.
126. Goertz, P., *How to escape from your sandbox and from your hardware host*, in *Available at: https://www.lesswrong.com/posts/TwH5jfkuvTatvAKEF/how-to-escape-from-your-sandbox-and-from-your-hardware-host*. July 31, 2015.

127. SethBling, *Jailbreaking Super Mario World to Install a Hex Editor & Mod Loader*. May 29, 2017: Available at: https://www.youtube.com/watch?v=Ixu8tn__91E.
128. Cooprocks123e, *Super Mario World Jailbreak Installer*. February 7, 2018: Available at: https://www.youtube.com/watch?v=lH7-Ua8CdSk.
129. SethBling, *SNES Code Injection -- Flappy Bird in SMW*. March 28, 2016: Available at: https://www.youtube.com/watch?v=hB6eY73sLV0.
130. Osgood, R., *Reprogramming Super Mario World from Inside the Game* in *Hackaday*. January 22, 2015: Available at: https://hackaday.com/2015/01/22/reprogramming-super-mario-world-from-inside-the-game/.
131. Burtt, G., *How an Ionizing Particle From Outer Space Helped a Mario Speedrunner Save Time*, in *The Gamer*. September 16, 2020: Available at: https://www.thegamer.com/how-ionizing-particle-outer-space-helped-super-mario-64-speedrunner-save-time/.
132. Anonymous, *SethBling*, in *Wikipedia*. Accessed on October 1, 2022: Available at: https://en.wikipedia.org/wiki/SethBling.
133. SethBling, *Route Notes: SNES Human Code Injection*. March 28, 2016: Available at: https://docs.google.com/document/d/1TJ6W7TI9fH3qXb2GrOqhtDAbVkbIHMvLusX1rTx9lHA.
134. Morgan, M.A., *Sepher ha-Razim: The Book of Mysteries*. Vol. 25. 2022: SBL Press.
135. Jonas, E. and K.P. Kording, *Could a neuroscientist understand a microprocessor?* PLoS computational biology, 2017. **13**(1): p. e1005268.
136. Gueron, S. and J.-P. Seifert. *On the impossibility of detecting virtual machine monitors*. in *IFIP International Information Security Conference*. 2009. Springer.
137. Demirbas, M., *Hacking the simulation*. April 1, 2019: Available at: http://muratbuffalo.blogspot.com/2019/04/hacking-simulation.html.
138. Canonico, L.B., *Escaping the Matrix: Plan B for Defeating the Simulation*. June 14, 2017: Available at: https://medium.com/@lorenzobarberiscanonico/escaping-the-matrix-plan-b-for-defeating-the-simulation-dd335988844.
139. Wasserman, R., *Paradoxes of time travel*. 2017: Oxford University Press.
140. Ford, A., *How to Escape the Matrix: Part 1*. January 21, 2015: Available at: https://hplusmagazine.com/2012/06/26/how-to-escape-the-matrix-part-1/.
141. Scharf, C.A., *Could We Force the Universe to Crash?* . 2020: Available at: https://www.scientificamerican.com/article/could-we-force-the-universe-to-crash/.
142. Torres, P., *Morality, foresight, and human flourishing: An introduction to existential risks*. 2017: Pitchstone Publishing (US&CA).
143. Benatar, D., *Better never to have been: The harm of coming into existence*. 2006: OUP Oxford.
144. Canonico, L.B., *Escaping the Matrix: Plan C for Defeating the Simulation*. July 30, 2017: Available at: https://medium.com/@lorenzobarberiscanonico/escaping-the-matrix-plan-c-for-defeating-the-simulation-e7d4926d1d57.
145. Chalmers, D., *Reality+: Virtual Worlds and the Problems of Philosophy*. 2022: W. W. Norton & Company.
146. Lanza, R., in *Psychology Today*. Decembre 22, 2021: Available at: https://www.psychologytoday.com/us/blog/biocentrism/202112/how-we-collectively-determine-reality.

147. Podolskiy, D., A.O. Barvinsky, and R. Lanza, *Parisi-Sourlas-like dimensional reduction of quantum gravity in the presence of observers.* Journal of Cosmology and Astroparticle Physics, 2021. **2021**(05): p. 048.
148. Wheeler, J.A., *The "past" and the "delayed-choice" double-slit experiment*, in *Mathematical foundations of quantum theory*. 1978, Elsevier. p. 9-48.
149. Orwell, G., *Nineteen eighty-four*. 2021: Hachette UK.
150. Maccone, L., *Quantum solution to the arrow-of-time dilemma.* Physical review letters, 2009. **103**(8): p. 080401.
151. Anderson, M.C. and B.J. Levy, *Suppressing unwanted memories.* Current Directions in Psychological Science, 2009. **18**(4): p. 189-194.
152. Nabavi, S., et al., *Engineering a memory with LTD and LTP.* Nature, 2014. **511**(7509): p. 348-352.
153. Ahire, J., *Reality is a Hypothesis*. Lulu. com.
154. Alexander, S., *The Hour I First Believed*. April 1, 2018: Available at: https://slatestarcodex.com/2018/04/01/the-hour-i-first-believed/.
155. Armstrong, S., *The AI in a Box Boxes You*, in *Less Wrong*. February 2, 2010: Available at: http://lesswrong.com/lw/1pz/the_ai_in_a_box_boxes_you/.
156. Bibeau-Delisle, A. and G. Brassard FRS, *Probability and consequences of living inside a computer simulation.* Proceedings of the Royal Society A, 2021. **477**(2247): p. 20200658.
157. Mullin, W.J., *Quantum weirdness*. 2017: Oxford University Press.
158. Turchin, A. and R. Yampolskiy, *Glitch in the Matrix: Urban Legend or Evidence of the Simulation?* 2019: Available at: https://philpapers.org/archive/TURGIT.docx.
159. Baclawski, K. *The observer effect*. in *2018 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*. 2018. IEEE.
160. Proietti, M., et al., *Experimental test of local observer independence.* Science advances, 2019. **5**(9).
161. Bong, K.-W., et al., *A strong no-go theorem on the Wigner's friend paradox.* Nature Physics, 2020. **16**(12): p. 1199-1205.
162. Lloyd, S., *Programming the universe: a quantum computer scientist takes on the cosmos*. 2007: Vintage.
163. Majot, A. and R. Yampolskiy, *Global Catastrophic Risk and Security Implications of Quantum Computers.* Futures, 2015.
164. Kim, Y.-H., et al., *Delayed "choice" quantum eraser.* Physical Review Letters, 2000. **84**(1): p. 1.
165. Schrödinger, E., *Die gegenwärtige Situation in der Quantenmechanik.* Naturwissenschaften, 1935. **23**(50): p. 844-849.
166. Cao, Y., et al., *Direct counterfactual communication via quantum Zeno effect.* Proceedings of the National Academy of Sciences, 2017. **114**(19): p. 4920-4924.
167. Gallego, M. and B. Dakić, *Macroscopically nonlocal quantum correlations.* Physical Review Letters, 2021. **127**(12): p. 120401.
168. Fein, Y.Y., et al., *Quantum superposition of molecules beyond 25 kDa.* Nature Physics, 2019. **15**(12): p. 1242-1245.
169. Krenn, M., et al., *Automated search for new quantum experiments.* Physical review letters, 2016. **116**(9): p. 090405.
170. Alexander, G., et al., *The sounds of science—a symphony for many instruments and voices.* Physica Scripta, 2020. **95**(6): p. 062501.

171. Wiseman, H.M., E.G. Cavalcanti, and E.G. Rieffel, *A" thoughtful" Local Friendliness no-go theorem: a prospective experiment with new assumptions to suit.* arXiv preprint arXiv:2209.08491, 2022.
172. Kirkpatrick, J., et al., *Pushing the frontiers of density functionals by solving the fractional electron problem.* Science, 2021. **374**(6573): p. 1385-1389.
173. Yampolskiy, R.V. *Analysis of types of self-improving software*. in *International Conference on Artificial General Intelligence*. 2015. Springer.
174. Anonymous, *Breaking out of a simulated world*. April 11, 2021: Available at: https://worldbuilding.stackexchange.com/questions/200532/breaking-out-of-a-simulated-world.
175. Bostrom, N., *Information Hazards: A Typology of Potential Harms From Knowledge.* Review of Contemporary Philosophy, 2011. **10**: p. 44-79.
176. Bruere, D., *The Simulation Argument — Jailbreak!* February 9, 2019: Available at: https://dirk-bruere.medium.com/the-simulation-argument-jailbreak-a61bd57d5bd7.
177. Baggili, I. and V. Behzadan, *Founding the domain of AI forensics.* arXiv preprint arXiv:1912.06497, 2019.
178. Schneider, J. and F. Breitinger, *AI Forensics: Did the artificial intelligence system do it? why?* arXiv preprint arXiv:2005.13635, 2020.
179. Ziesche, S. and R. Yampolskiy, *Designometry–Formalization of Artifacts and Methods*. Available at: https://philarchive.org/archive/ZIEDF.
180. Ziesche, S. and R. Yampolskiy, *Towards AI welfare science and policies.* Big Data and Cognitive Computing, 2018. **3**(1): p. 2.
181. Yampolskiy, R.V., *AI Personhood: Rights and Laws*, in *Machine Law, Ethics, and Morality in the Age of Artificial Intelligence*. 2021, IGI Global. p. 1-11.
182. Moravec, H., *Simulation, consciousness, existence.* Intercommunication, 1999. **28**.
183. Yampolskiy, R.V., *Leakproofing Singularity - Artificial Intelligence Confinement Problem.* Journal of Consciousness Studies (JCS), 2012. **19(1-2)**: p. 194–214.
184. Turchin, A., *Catching Treacherous Turn: A Model of the Multilevel AI Boxing*. 2021: Available at: https://www.researchgate.net/profile/Alexey-Turchin/publication/352569372_Catching_Treacherous_Turn_A_Model_of_the_Multilevel_AI_Boxing.
185. Babcock, J., J. Kramar, and R. Yampolskiy, *The AGI Containment Problem*, in *The Ninth Conference on Artificial General Intelligence (AGI2015)*. July 16-19, 2016: NYC, USA.
186. Babcock, J., J. Kramár, and R.V. Yampolskiy, *Guidelines for artificial intelligence containment*, in *Next-Generation Ethics: Engineering a Better Society*, A.E. Abbas, Editor. 2019. p. 90-112.
187. Yudkowsky, E.S., *The AI-Box Experiment*. 2002: Available at: http://yudkowsky.net/singularity/aibox.
188. Armstrong, S. and R.V. Yampolskiy, *Security solutions for intelligent and complex systems*, in *Security Solutions for Hyperconnectivity and the Internet of Things*. 2017, IGI Global. p. 37-88.
189. Alfonseca, M., et al., *Superintelligence cannot be contained: Lessons from Computability Theory.* Journal of Artificial Intelligence Research, 2021. **70**: p. 65-76.
190. Yampolskiy, R. *On the Differences between Human and Machine Intelligence*. in *AISafety@ IJCAI*. 2021.

191. Kastrenakes, J., *Facebook is spending at least $10 billion this year on its metaverse division*. October 25, 2021: Available at: https://www.theverge.com/2021/10/25/22745381/facebook-reality-labs-10-billion-metaverse.
192. Bostrom, N. and E. Yudkowsky, *The Ethics of Artificial Intelligence*, in *Cambridge Handbook of Artificial Intelligence. Cambridge University Press*, W. Ramsey and K. Frankish, Editors. 2011: Available at http://www.nickbostrom.com/ethics/artificial-intelligence.pdf.
193. Astin, J.A., E. Harkness, and E. Ernst, *The efficacy of "Distant Healing" a systematic review of randomized trials.* Annals of internal medicine, 2000. **132**(11): p. 903-910.
194. Sleiman, M.D., A.P. Lauf, and R. Yampolskiy. *Bitcoin Message: Data Insertion on a Proof-of-Work Cryptocurrency System*. in *2015 International Conference on Cyberworlds (CW)*. 2015. IEEE.
195. D, R., *The Opt-Out Clause*. November 3, 2021: Available at: https://www.lesswrong.com/posts/vdzEpiYX4aRqtpPSt/the-opt-out-clause.
196. Gribbin, J., *Are we living in a designer universe?* . August 31, 2010: Available at: https://www.telegraph.co.uk/news/science/space/7972538/Are-we-living-in-a-designer-universe.html.
197. Anonymous, *A series on different ways to escape the simulation*. 2017: Available at: https://www.reddit.com/r/AWLIAS/comments/6qi63u/a_series_on_different_ways_to_escape_the/.
198. Virk, R., *The Simulation Hypothesis: An MIT Computer Scientist Shows Why AI, Quantum Physics, and Eastern Mystics All Agree We Are in a Video Game*. 2019: Bayview Books, LLC.
199. Adamson, R., *Hacking the Universe*. November 4, 2018: Available at: https://hackernoon.com/hacking-the-universe-5b763985dc7b.
200. Yampolskiy, R.V., *Efficiency Theory: a Unifying Theory for Information, Computation and Intelligence.* Journal of Discrete Mathematical Sciences & Cryptography, 2013. **16(4-5)**: p. 259-277.
201. Gates, J., *Symbols of power: Adinkras and the nature of reality.* Physics World, 2010. **23**(6).
202. Turchin, A. and R. Yampolskiy, *Types of Boltzmann brains*. 2019: Available at: https://philarchive.org/archive/TURTOB-2.
203. Heylighen, F., *Brain in a vat cannot break out.* Journal of Consciousness Studies, 2012. **19**(1-2): p. 1-2.
204. Turchin, A., *Multilevel Strategy for Immortality: Plan A–Fighting Aging, Plan B–Cryonics, Plan C–Digital Immortality, Plan D–Big World Immortality*. Available at: https://philpapers.org/go.pl?id=TURMSF-2&proxyId=&u=https%3A%2F%2Fphilpapers.org%2Farchive%2FTURMSF-2.docx.
205. Ettinger, R.C., *The prospect of immortality*. Vol. 177. 1964: Doubleday New York.