Intelligent Information Management, 2024, 16, 121-146
https://www.scirp.org/journal/iim
ISSN Online: 2160-5920
ISSN Print: 2160-5912

Tackling the Existential Threats from Quantum Computers and AI

Fazal Raheman

Blockchain 5.0 Ltd., Tallinn, Estonia
Email: drfazal@bc5.eu

How to cite this paper: Raheman, F. (2024) Tackling the Existential Threats from Quantum Computers and AI. *Intelligent Information Management*, **16**, 121-146.
https://doi.org/10.4236/iim.2024.163008

Received: April 10, 2024
Accepted: May 28, 2024
Published: May 31, 2024

Abstract

Although AI and quantum computing (QC) are fast emerging as key enablers of the future Internet, experts believe they pose an existential threat to humanity. Responding to the frenzied release of ChatGPT/GPT-4, thousands of alarmed tech leaders recently signed an open letter to pause AI research to prepare for the catastrophic threats to humanity from uncontrolled AGI (Artificial General Intelligence). Perceived as an "epistemological nightmare", AGI is believed to be on the anvil with GPT-5. Two computing rules appear responsible for these risks. 1) Mandatory third-party permissions that allow computers to run applications at the expense of introducing vulnerabilities. 2) The Halting Problem of Turing-complete AI programming languages potentially renders AGI unstoppable. The double whammy of these inherent weaknesses remains invincible under the legacy systems. A recent cybersecurity breakthrough shows that banning all permissions reduces the computer attack surface to zero, delivering a new zero vulnerability computing (ZVC) paradigm. Deploying ZVC and blockchain, this paper formulates and supports a hypothesis: "Safe, secure, ethical, controllable AGI/QC is possible by conquering the two unassailable rules of computability." Pursued by a European consortium, testing/proving the proposed hypothesis will have a groundbreaking impact on the future digital infrastructure when AGI/QC starts powering the 75 billion internet devices by 2025.

Keywords

Ethical AI, Quantum Computers, Existential Threat, Computer Vulnerabilities, Halting Problem, AGI

1. Introduction

The existential threat to humanity and the "epistemological nightmare" from Artificial intelligence (AI) is a matter of the moment [1]. So is Quantum Com-

puting [2]. Both are rapidly evolving as potential tools of destruction that adversaries can potentially exploit [3].

For a very long time, AI has been a subject of interest among fiction writers and sci-fi communities. In recent decades, AI has demonstrated the potential to step out of fiction and soon become a reality, replicating human-level intelligence. Human intelligence is as complex as human behavior and cognition. It can be defined in multiple different ways. A machine that can understand or learn intellectual undertakings to the capacity that humans can is characterized as artificial general intelligence (AGI). Machine learning (ML) algorithms have been developed to build a wide range of specialized AI applications that are getting better than humans at specific tasks. Artificial neural networks are being developed to mimic the way human brain works. As the evidence that AI can execute tasks better and cost-effectively accumulates, every industry, directly or indirectly deploying computers, embraces AI. The rapid industrialization of AI is increasingly becoming a cause of concern because of its vulnerabilities and misuse by bad actors [1] [3].

Quantum computing (QC) is a multidisciplinary field that combines aspects of computer science, physics, and mathematics, utilizing quantum mechanics to solve complex problems much faster than classical computers. Because of its extraordinary computing speed, QC can easily decrypt today's encryption schemes to break the Internet [4]. Since access to the Internet has become as important a need as necessities of life, such as electricity, shelter, potable water, smart phones, communication, etc., life without the Internet is unimaginable. In almost everything we do today, we use the Internet. Therefore, any security risk to the Internet is also considered an existential risk to humanity and needs to be mitigated with some urgency [2] [4].

Whether AI benefits from QC is no longer a question now [5] as Quantum Machine Learning has become a dedicated area of research [6]. Intricately intertwined, QC and AI are changing the world at a pace that's scaring the conservatives [1] [2] [3] [4]. Recent advances in large language models (LLMs) that use deep learning (DL) techniques and large data sets to comprehend, summarize, generate, and predict new content are revolutionizing the AI space. GPT (Generative Pre-trained Transformer) is one of the largest LLMs developed by OpenAI. GPT-3.5 was launched as ChatGPT on November 30, 2022, generating an explosion of interest globally [7]. With over 100 million active users in just the first two months, ChatGPT set a world record for the fastest-growing user base of any application in history [8].On March 14, 2023, OpenAI released the latest generation of the large-scale multimodal language model, GPT-4, as ChatGPT Plus [9]. The release of GPT-4 has caused an uproar worldwide on speculation that the next version of GPT (GPT-5) may be AGI. Experts believe the early experiments with GPT4 already show early signs of AGI [10] and that GPT-5 itself may be AGI [11]. ChatGPT *5* will make it feel like you communicate with someone human rather than a machine. Some experts believe that GPT-5 has the potential to

achieve super intelligence. They argue that GPT-5's ability to learn and adapt, combined with its vast dataset of text and code, could allow it to eventually surpass human intelligence. Perhaps that is the reason why thousands of AI experts and stakeholders signed a petition to pause further GPT-5 development for at least six months [12] [13], and within weeks, followed up with another "Statement on AI Risk" asserting: "*Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war* [14]." These proclamations have sparked broad discussions and controversies across the world. Prominent academicians and journal editorials warn that "*if we do not control our own civilization, we have no say in whether we continue to exist* [15] [16]". With strong proponents on both sides of the debate, whether AI's existential threat is real or just fearmongering [17] is impossible to judge. The pause-AI call may not be unanimous, but opinions on AI risk appear to be unanimous, as reflected by the opinion of a non-signatory expert.

"AI is out of the bag—it cannot and will not be stopped or paused, for better or for worse. Our best bet is to develop proactive before-the-event policies, risk management frameworks, and safeguards, along with aggressive and accelerated development of the compliant side of AI developers to ensure that the 'good' side stays ahead [18]*."*

1.1. Will AI Go Rogue?

Whether AI will go rogue in 2023, 18 - 53% (27% global average) of 24,471 adult respondents of a recent survey across 34 countries answered YES (**Figure 1**) [19]. If the perception that a rogue AI program will cause problems worldwide within this year is so high, then its future likelihood cannot be just dismissed. When the general perception regarding the dangers of AI is unprecedented, and when thousands of the world's top tech luminaries sign not one but two open

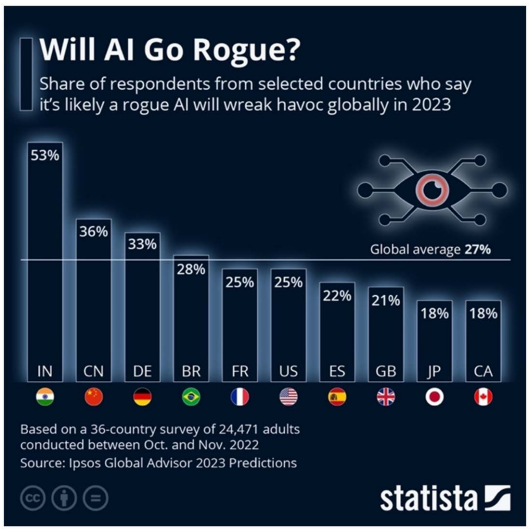

Figure 1. A survey of 24,471 adults from 36 countries.
Source: https://www.statista.com/.

letters within weeks, and the world's first AI regulation gets closer to being legislated [20], the concerns of "existential threat from AI" needs to be addressed with some urgency. As the debate for and against the impending threats intensifies, the need to secure our digital infrastructures is real and immediate. Human accessibility to immensely powerful tools like superhuman AI and QC can never be without apocalyptic risk to humanity. Unless technological advances are not moderated and democratized, the risk of their misuse can never be ruled out.

Surprisingly, there is so much chatter about AI's existential risk but very little or virtually nothing about why the AI threat is technically unassailable. The problem can only be adequately addressed if the exact cause of the problem is identified. Before we embark on that journey, it is essential to understand what safe, excellent, or responsible AI means.

1.2. What Do Robustness, Resilience, Ethics, and Security Mean for AI?

We must first understand that AI is not a monolithic term when considering safe and responsible AI. AI is a phenomenon that needs to be seen in a more nuanced way through the lens of its evolutionary stages comprising of ANI (artificial narrow intelligence), AGI (artificial general intelligence), and ASI (artificial super intelligence) [21]. Its robustness, resilience, and security must be evaluated at these evolutionary stages to assess its full potential and risks. Today, we have already achieved ANI, and as we move towards AGI and eventually ASI, the burden on the parameters for robustness, resilience, and security gets heavier. Since the future of AI is predicted to be in quantum computing (QC) [22] [23], additional vulnerabilities of QC as an amplifier of the existential risk [24] cannot be ignored when defining those parameters. All those considerations are considered in articulating the problem statement for structuring this perspective. The terms AI and AGI are interchangeably used throughout this article to imply risky aspects of AI.

1.3. When Will Q-Day Arrive?

In 2016, NIST (National Institute of Standards and Technology) published a report on the rising threat to encrypted Internet data by quantum computers and the catastrophic impact that it would have on the integrity of the global IT infrastructure [4]. Following the NIST report, experts have warned of the apocalyptic Q-Day when QC will have enough computing power to decrypt state-of-the-art encryptions and break the Internet [25]. The exponential growth in QC has opened up the possibility of performing attacks based on Shor's and Grover's algorithms that threaten the PKI and hash functions in the near future [26]. QC may still be far from becoming mainstream, but there is a big push to bring it to the mainstream soon [27]. In a recent survey, 74% of IT professionals believe QC with sufficient Qubits to break legacy encryption algorithms will arrive in five years [28]. According to a Y2Q (years to quantum) clock launched by the Cloud Security Alliance last year, the Q-Day may be just under seven years before

Figure 2. Countdown to Q-Day (Y2Q). Credit: Cloud Security Alliance.

QC can crack current encryption (**Figure 2**) [29]. The timeline estimates projected for AGI, QC, and when smart cities become a norm more or less culminate around 2030 [30].

2. Problem Statement & the Hypothesis

Whether AGI or QC, their integration into our smart cities and lives is imminent. A UNESCO report predicts smart cities will shape our societies by 2030 [31]. This paper investigates a fundamental research question that essentially transcends all information technology-related fields directly or indirectly impacted by AI, QC, and cybersecurity. AI and QC are two frontiers of research in computer science that meet in the brand-new field of quantum artificial intelligence [32], and cybersecurity is the backbone of any successful digitalization of modern society [33]. AI and QCare very active fields of computer research with an overwhelming speed of new developments. The processing power to create the human brain is enormous, but QC might be our gateway to successfully creating AGI in the future [34]. Like AI, QC present sits own existential risk via its ability to break the Internet with its decryption capacity [25] [35] [36]. Cybersecurity is the common denominator in the success of both these fields. Hence, whether it is the future of AI or QC, cybersecurity remains at the epicenter of its real-world implementation. With that backdrop to the problem that AI/QC faces today and in the near future following hypothesis is formulated:

Safe, secure, ethical, and controllable AGI/QC is possible by conquering the two unassailable rules of computability with Collective Artificial Super Intelligence (CASI).

The hypothesis may appear ambitious, but it is grounded in solid science and supported with empirical evidence. All AI systems are essentially computer programs, and as such, they inherit the limitations imposed by the rules of computing. A computer's greatest strength is also its greatest weakness. What makes them so powerful, widespread, and valuable also places a fundamental limit on what they can do and the problems they can solve. This limit cannot be broken in the prior art, even with supercomputers or quantum computers. It is built into the very nature of computation. What is this intriguing paradoxical quality? Well, there are not one but two paradoxes (**Figure 3**). The existential threats to humanity from AI and QC are routed in those qualities of computation that render computing systems vulnerable and cannot be circumvented in the current state-of-the-art. These rules of computation are:

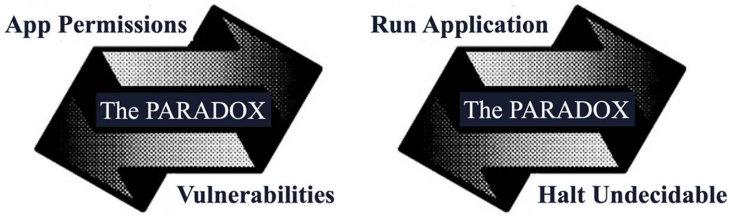

Figure 3. The permissions paradox & the halting paradox.

1) Third-party permissions are mandatory in building present-day computer hardware and software [36] [37] [38] [39]. These permissions make it possible to run a wide range of applications from third-party vendors but are also responsible for the vulnerabilities that bad actors often exploit. That is why all legacy computers remain vulnerable [38].

2) As a basic computability rule, deciding whether a specific Turing machine should halt or run infinitely is undecidable [40]. Termed as "*the Halting Problem*", this phenomenon renders AI/AGI unstoppable and uncontainable if it goes rogue [41]. Turing proved that "*a general algorithm to solve the halting problem for all possible program-input pairs cannot exist* [42] [43]".

Both paradoxes remain invincible under legacy computing systems and remain the principal cause of AI/AGI's existential threat to humanity. In other words, a functional computing system completely free from vulnerabilities and a Turing-complete algorithm that can guarantee the halt of an adverse algorithm if it continues to run in a loop indefinitely cannot exist. To resolve these paradoxes following research questions can be formulated:

i) Can ZVC (Zero Vulnerability Computing) provide autonomous and seamless security to AI and Quantum Computing?

ii) Will the integration of blockchain render AI controllable?

Having articulated the research questions, reviewing the research methodology (Section 3) and the current state-of-the-art approach to AI safety, security, and containment is essential for building the hypothesis (Section 4). This will help place the evidence in support of the CASI hypothesis in the proper perspective. So, the next sections discuss the state-of-the-art, followed by section 5 on the evidence supporting the hypothesis that goes beyond the state-of-the-art. Section 6 highlights the prospects of proving the hypothesis and limitations of the proposed research, and finally, section 7 summarizes and discusses the conclusions of this research.

3. Research Methodology

This is hypothesis-generating research designed to generate and formulate a new research question. It is not a hypothesis-testing or hypothesis-proving research designed to empirically answer a known research question [44]. Although the methods of less rigorous hypothesis-generating research do not replace or undermine more rigorous hypothesis-testing or hypothesis-proving research me-

thodologies, they play an important role in building new paradigms that lay the foundation for discoveries. Except for a few rare serendipitous inventions, almost all discoveries the world has ever seen begin with a **HYPOTHESIS**. Whether a hypothesis is eventually proven or disproven, it never loses its importance as the beginning of a journey to new knowledge. Historically, hypothesis-generating research has facilitated inventions that may not have been possible otherwise [45]. The computability rules, viz. mandatory third-party permissions [36] [37] [38] [39], and the halting problem of turing complete programming languages [40] [41] [42] [43], which this paper acknowledges in formulating the CASI hypothesis in the previous section, have existed since modern computers came into existence. However, in peer-reviewed literature, they are assumed to be inseparable from the computing systems and scarcely studied in the cybersecurity or AI context except in a handful of studies [36]-[43]. A detailed literature review is presented herein to formulate and support the hypothesis.

4. Review of Literature: Foolproof Security & Controllability of AI Impossible?

The seven requirements for trustworthy AI articulated by AI HLEG that render AI lawful, ethical, and robust must be secure and controllable (**Figure 4**). Saghiri *et al.* surveyed challenges that AI faces today and concluded that in the era of super intelligence or AGI, the ML agents would be difficult to control for humans [46]. They identified 28 challenges that AI needs to address. In the peer-reviewed literature, at least two challenges are unassailable and essentially the cause of classifying AI as an existential risk to humanity. As illustrated in **Figure 4**, security and controllability are the two challenges that constitute the two foundational pillars on which trustworthy AI/AGI of the CASI hypothesis is built.

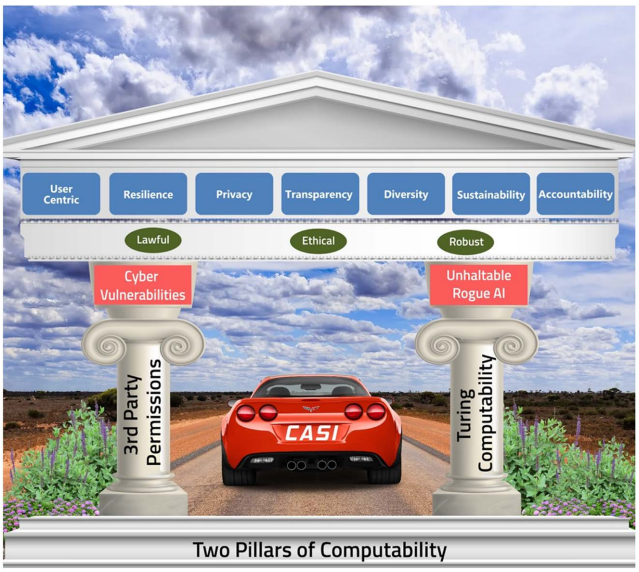

Figure 4. Overcoming the mandates of computability to build lawful, ethical, robust AI system.

As identified in the preceding section, these foundational pillars represent two mandatory rules of computing that any legacy AI system must comply with, viz. i) third-party permissions and ii) Turing computability (**Figure 4**). As discussed, legacy computers and AI systems can only be built in compliance with these rules. Put another way, these two mandatory computing rules render AI/AGI vulnerable to adversarial attacks and cannot guarantee foolproof security from adversarial control of AGI due to hacking or unstoppability if it goes into rogue hands.

The controllability of AI/AGI has four types: explicit, implicit, delegated, and aligned, and it gets more severe by increasing the autonomy of AI-based agents [46] [47] [48]. Consequently, properly balancing security with usability is a major concern when designing any AI containment strategy. The tradeoff between security and usability is a tough question without clear answers. In extreme cases, the most secure method could render AI useless, defeating the whole purpose of building AI. Ignoring the AI's core uncontrollability problem, Babcock *et al.* discuss AGI containment with different tradeoffs between usability and security [49]. However, because of the assumed capabilities of future AGI/ASI, we cannot rule out the possibility of machines being uncontrollable in some situations because of their halting problem [40] [41] [42] [43].

4.1. Why Is Foolproof AGI Cybersecurity Impossible in Legacy Systems?

AI systems are computer programs. As such, they remain subject to the basic rules of computability theory. Third-party permissions are mandatory in building present-day computers and AI algorithms. However, these permissions are also responsible for the vulnerabilities that bad actors can exploit [36] [37] [38] [39]. It is undeniable that all computers are vulnerable [49], and therefore, all cybersecurity implementations are policy-based and cannot be secured by default or by design. Standard security measures for ML models [50] include (i) access control, (ii) system monitoring, and (iii) audit trail.

The principal reason all computers are vulnerable, and no computer is without an attack surface, is because computer hardware and software cannot be built without granting third-party permissions that software vendors can use to develop applications that make computers work. Bad actors often abuse these permissions by creating attack vectors that render computers vulnerable to malware. Without such permissions, computers will be virtually useless as none of the diverse range of software applications we depend on will work. So, a paradoxical catch-22 situation makes these third-party permissions a necessary evil that remains unassailable in the prior art [36] [37] [38] [39].

The traditional attack surface results from third-party permissions that all computers mandate for running third-party applications [36] [37] [38] [39]. The advent of AI must deal with additional vulnerabilities associated explicitly with machine learning (ML) that create an ML attack surface [51]. The ML at-

tack surface results from training data sets, which attackers can manipulate well before model deployment time. Such attack vectors, which do not exist in conventional software, include adversarial reprogramming, data poisoning, malicious input, or stealing information by a probe [52] (**Figure 5(a)**). Discussed in more detail in the next section, **Figure 5(a)** illustrates all these vulnerabilities resulting from the traditional as well as ML attack surface in a self-explanatory graphic illustration adapted from Isaac & Reno [53]. Cyber attackers use threat vectors to target the vulnerable attack surface. NIST defines a threat as *"The potential for a threat source to exploit (intentionally) or trigger (accidentally) a specific vulnerability* [54]". The literature describes several types of threat modeling approaches [55]. STRIDE is the most mature and widely used strategy [56] that Mauri & Damiani recently adapted for threat modeling in the AI domain by using an *asset-centered* methodology for identifying threats to ML-based systems [57]. Their *STRIDE-AI* strategy for assessing the security of AI-ML-based systems identifies six asset classes in the AI ecosystem (**Figure 6**) [58]. They argue that their STRIDE-AI extension to the original STRIDE provides an *ML-specific*, *security property-driven* approach to threat detection, which can also guide in selecting the security controls needed to alleviate the identified threats. In a special report on AI Cybersecurity, the European Union Agency for Cybersecurity (ENISA) also considers STRIDE a promising starting point for AI threat modeling [59].

The STRIDE strategy is an acronym for six threat vectors that hackers commonly use to attack any computing resource connected to a network. They are Spoofing, Tampering, Repudiation, information disclosure, Denial-of-Service, and Elevation of Privilege [58], as illustrated in the following **Table 1**.

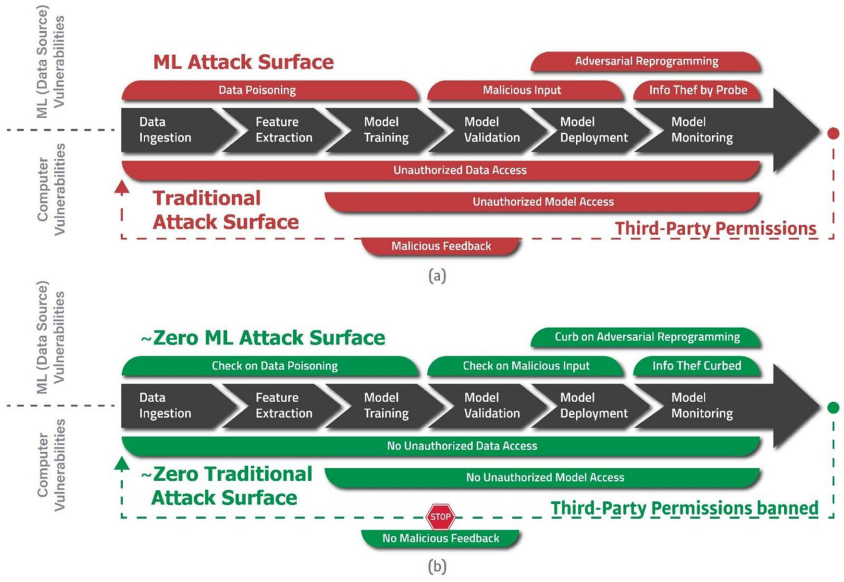

Figure 5. Vulnerabilities and Threats in (a) Legacy ML Systems vs. (b) CASI adopted from Isaac & Reno, https://arxiv.org/pdf/2304.11087.pdf.

As illustrated in Figure 6, except in some instances of DoS/DDoS attacks on AI assets, the remaining five require subscription-based authentication for accessing the AI assets. Therefore, the only way a cyber attacker can generate a threat vector in these cases is by gaining access to the AI assets. In other words, except for some types of DoS/DDoS attacks, a hacker will always need to access an ML agent as a legitimate subscriber or an unauthorized intruder using an attack vector to breach the ML attack surface. As already explained, the traditional attack surface is essentially a consequence of third-party permissions that all computers mandate for running third-party applications [36] [37] [38] [39]. Therefore, as in traditional security breaches, the ML attack surface also depends

Table 1. Six threats of STRIDE.

Threat	Description
Spoofing Identity	An attacker poses as an authorized user by taking or faking an identity of another person.
Tampering with Data	An attacker modifies some information in the system by changing a data item.
Repudiation	An attacker deletes a transaction to cover up and deny his intrusion into the system.
Information Disclosure	Personal user data is stolen and sold to a competitor with an intent to make profit.
Denial of Service (DoS)	An attacker exhausts network resources to make it inaccessible to its intended users.
Elevation of Privilege (EoP)	*An attacker, instead of spoofing identity, just elevates his own security level to an administrator.*

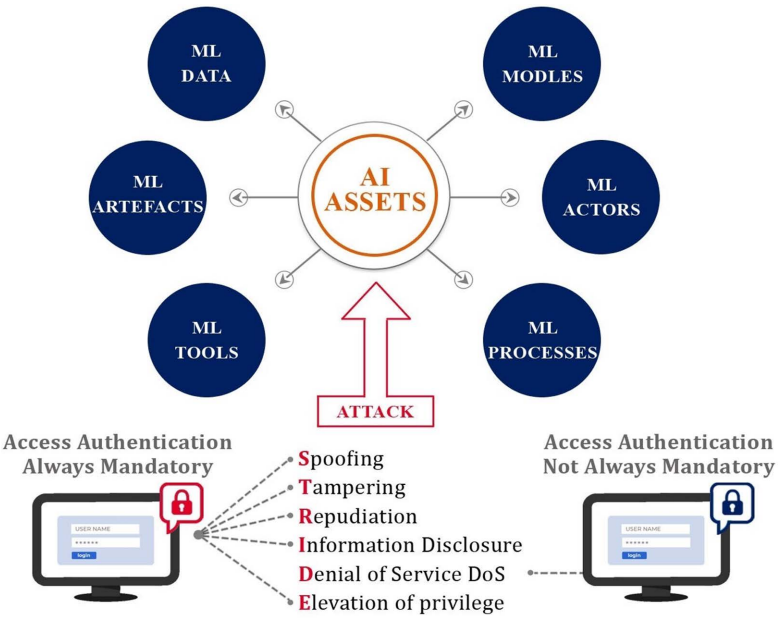

Figure 6. Holistic STRIDE-AI Strategy to secure ML Assets. Adapted from Mauri and Damiani, doi: 10.1109/CSR51186.2021.9527917 [57].

on unauthorized access to ML assets. Such unauthorized access is only possible because all legacy systems integrate third-party permissions [36] [37] [38] [39], which bad actors abuse by deploying different strategies to attack the ML models of AI's neural network.

4.2. AI's Unassailable Halting Problem Renders AI Unstoppable

Another basic rule of the computability theory presents a pivotal limitation to the AI algorithms programmed in Turing-complete programming languages. It is impossible to write a program in Turing-complete language that can examine any other program and tell, in every case, if it will terminate or get into a closed loop when it is run [40] [41] [42] [43] [60]. Termed the **halting problem**, it is *one of the most philosophically important theorems of the "theory of computation*" and is unsolvable [61]. The undecidability of the halting problem has an immediate practical bearing on all software development, particularly in AI development. Widely regarded as the canonical undecidable problem [62], the halting problem [63] and controllability [64] of a Turing complete program are impossible to solve [65]. Therefore, many AI-based systems that inherit the problem might not be manageable. "It is simply not possible for computers to catch the halting problem. Humans will always be a part of it [65] [66]." Alfonseca *et al.* argue that total containment of super intelligence is principally impossible due to the fundamental limits inherent to the theory of computing itself [41]. The halting problem is a decision problem about the properties of computer programs on a fixed Turing-complete model of computation. The problem is to determine, given a program and an input to the program, whether the program will eventually halt or run indefinitely when executed with that input. In simple terms, "halting problem" refers to the impossibility of determining whether an arbitrary computer program will finish running or continue to run forever (**Figure 7(a)**). Turing proved that "*a general algorithm to solve the halting problem for all possible program-input pairs cannot exist* [67] [68]". This

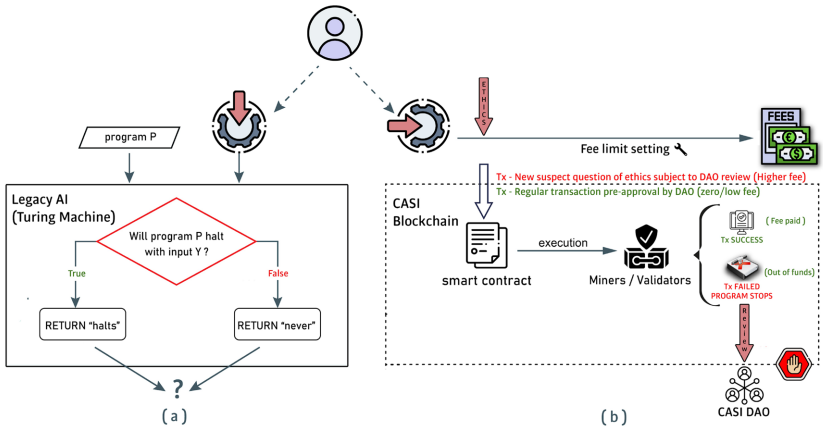

Figure 7. Undecidable Halting Problem of (a) Legacy AI vs. (b) DAO Controlled CASI system.

means that the only way to stop a globally rogue AI would be something that must go wrong with the Internet infrastructure itself, which is highly unlikely. In simple terms, once a self-referential loop is encountered, the program that generated it cannot "step out" of itself. An outside source needs to address the problem, which in the case of a computer running AI algorithms is eventually a human operator.

4.3. The Ethical Dilemma with AGI/ASI

Can an AI system developer, implementer, or regulator, utilizing any known procedure, predetermine whether an AI system consistently delivers output that complies with ethical norms? The answer is "NO". No algorithm in any of the Turing-complete languages can reliably do so for all AI systems all the time for any given input. Although, for some AI systems running some input, ethical control may be possible, but not always, leaving AGI/ASI unrestrained by any ethical norms [68]. So, ethical compliance also remains undecidable and consequently unsolvable.

Five principles guide the ethical governance of AI in society: i) beneficence (promote human well-being), ii) non-maleficence (do no harm), iii) autonomy (preserve human freedom), iv) justice (operate with fairness), and v) explicability (output explainable results) [69]. However, AGI and ASI pose a fundamentally different problem than those typically studied by Asimov under the banner of "robot ethics" [70]. This is because AGI and ASI are multi-faceted and, therefore, capable of mobilizing a diversity of resources to achieve potentially incomprehensible objectives to humans, let alone controllable [41]. As argued in the preceding section, the Halting Problem introduces subjectivity of decision output at all levels [71], making run-time implementation of algorithms with principles of ethics impossible.

5. Beyond State-of-the-Art

A solid cybersecurity infrastructure is the most crucial shield against data poisoning or other AI breaches [72]. However, legacy systems cannot be totally free from vulnerabilities. They remain vulnerable because of their mandatory permissions to use the third-party codes built into their architecture [36] [37] [38] [39]. Put another way, all legacy computing systems can run any third-party code/algorithm irrespective of whether it is installed by legitimate means or injected by a bad actor. This means a legacy system always leaves some attack surface that an adversary can disguise as a legitimate system user to gain access and exploit AI's ML algorithms in several ways.

5.1. Zero Vulnerability Computing (ZVC)

A breakthrough in cybersecurity provides an interface for running all third-party applications without granting them any permission to install on the computer [36] [37] [38] [39]. ZVC is a new award-winning computing paradigm [73] that

bans all third-party permissions to eradicate computer vulnerabilities and reduce the attack surface to zero and is defined as follows:

"ZVC is a cybersecurity paradigm that proposes a new zero attack surface computer architecture that restricts all third-party applications exclusively to a web interface only, declining permissions for any utilization of computing resources by any non-native program and creates a switchable in-computer offline storage for securing sensitive data at the user's behest [36]*."*

The empirical evidence unequivocally supports the ~zero attack surface proposition and establishes that ZVC is inherently resistant to threats from QC because its security mechanism is encryption-agnostic and not cryptography-dependent [36] [37] [38] [39]. If a technology can secure a connected computer without the need for user-facing access authorizing encryption algorithms, it will automatically make the connected device quantum-safe [38]. Moreover, all legal security strategies, including "Zero Trust", are policy-based and implemented by humans and cannot be autonomous and seamless [39]. This imposes a limitation on implementing these strategies in AGI, as AGI implementation in Generative AI, robotics, or autonomous mobility is expected to be autonomous [21]. ZVC runs seamlessly and autonomously without human intervention or monitoring [36] [37] [38] [39].

As stated earlier in the previous section, Isaac & Reno [53] recently summarized the vulnerabilities of the traditional and ML attack surface in a self-explanatory graphic illustration. Based on an adaptation of their illustration, we elucidate how CASI deploys ZVC to eliminate or mitigate these attack surfaces (**Figure 5(b)**). By banning all third-party permissions, ZVC obliterates the traditional attack surface. The advent of AI has introduced another type of attack surface resulting from bad actors deploying different strategies to attack machine learning (ML) models of AI's neural network. While traditional attack surface is essentially an outcome of third-party permissions, except DoS (denial of service) attacks (**Figure 6**), a good majority of ML attacks are also permission dependent (See **section 3.1** and **Figure 6**). As illustrated in **Figure 5(b)** the empirical evidence [36] [37] [38] [39] suggests that ZVC can also potentially eliminate ML Attack Surface as it is also primarily permission-based. This is explained with some clarity in **Figure 5(b)**, how ML attack surface is created, and how ZVC can deal with it.

5.2. How Does CASI Solve the Halting Problem to Stop Rogue AI?

As we have seen, banning all third-party permissions with ZVC obliterates the traditional and ML attack surface to keep the bad actors from breaching a responsible AI (**sec 4.1**). As far as much of the traditional as well as ML attack surface originates from the rogue elements' accessibility to ML models, data poisoning, adversarial reprogramming, malicious input, or stealing information by a probe (**Figure 5**), the CASI ecosystem can impede unauthorized access from legacy computing devices, tackling all known AI attack vectors. However, if

a bad actor succeeds in taking control of CASI by some mechanism presently unknown, the halting problem may still render AI unstoppable [41]. As a standard computability rule, in any of the Turing complete AI programming languages, no algorithm can determine if a program would halt and not run into an infinite loop [40] [41] [42] [43] [60]-[68]. Therefore, there is no way to exercise any control over the adverse actions of AI if it goes into rogue hands. Using a blockchain-based AI governance strategy, CASI builds an indirect defense against the unsolvable Halting problem to stop rogue AI. Blockchain is a continuously growing data ledger. Machine learning algorithms can be trained on smart blockchain contracts [74] to produce trusted models for reliable prediction [75]. A smart contract is made between all the AI stakeholders and deployed on the blockchain. Smart contracts are self-executing agreements encoded on a blockchain with their terms directly inscribed in code that will automatically execute when predetermined conditions are met [74] [75]. Smart contracts offer transparent, tamperproof, and cost-effective alternatives to traditional contracts. Peer-reviewed literature presents many examples of Blockchain-enabled AI systems in diverse use case settings, such as ensuring accountability and quality control in zero defect manufacturing [76], enhancing edge intelligence in IoT networks [77], for automatic learning in big data-based digital gaming [78], for improving cybersecurity [79], in healthcare [80], and many more [81]. However, in the literature, one cannot find evidence of using blockchain smart contracts to solve AI's halting problem. The CASI framework provides an indirect solution to the halting problem by deploying any one of the following two approaches:

a) Obstructing execution of any new unethical decision by smart contract transaction fee restriction

Because AI programs are Turing complete, the halting problem applies, and a single execution of a rogue AI could run forever [60]-[68]. To prevent this, CASI uses a specially designed blockchain that assigns certain unethical or rogue decision-making to smart contracts that can only be executed autonomously if a sufficient fee (gas) is available in the system wallet payable to the miner/validator. Miners or validators of blockchain transactions must spend resources such as computing power or electricity to validate and record each transaction on the blockchain [82]. Such costs are recovered as transaction fees, generally calculated based on the transaction size in bytes and the current network congestion. A miner/validator will terminate the script if it runs out of funds (**Figure 7(b)**). Thus, blockchain indirectly addresses the halting problem by introducing the concept of gas (transaction fee) [82]. All ML actions are divided into two types of transactions,

i) routine, no-fee ML actions pre-approved by the DAO (decentralized autonomous organization) that governs the blockchain;

ii) All new suspect ML actions require fee-based smart contract authentication, wherein the DAO controls such fee remittance.

As illustrated in **Figure 7**, all routine transactions get executed without any

restrictions. However, any new suspected unethical action will only execute if fee restrictions pause the AI engine until such actions get reviewed by the DAO or DAO's ethical committee. By making the CASI wallet multi-sig (requiring multiple unique signatures) under the control of the democratically elected CASI DAO members, the execution of the smart contract is guaranteed to halt the Turing-complete algorithm destined for infinite loops [83]. Thus, CASI indirectly solves the unassailable Halting Problem and can prevent AGI/ASI from going rogue.

b) Coding smart contract using a non-Turing compete language

Although it is a standard practice to use Turing complete programming language to code the conventional smart contract, recent evidence suggests that smart contracts can also be efficiently coded using a non-Turing complete language [84]. A non-Turing complete language such as Vyper does not face the halting problem, and smart contracts coded in Vyper are more efficient in terms of performance speed, storage, and eliminating certain classes of bugs [85]. This means a CASI smart contract coded in a non-Turing language can automatically stop anytime the ML detects an unethical anti-human action without resorting to the indirect method of stopping smart contract execution utilizing fee restriction. Non-Turing-complete smart contracts allow easier auditing due to the lower code complexity since they do not support recursion or complex loops [84] [85]. This will also decrease the possibility of implementing defects since the code will be more straightforward to review. Executing simpler programs results in better performance and prevents congestion, often caused by Turing-complete smart contracts that use much storage [86]. Our current development focuses on studying the merits of both approaches regarding ease of implementation and effectiveness in stopping rogue AI/AGI.

5.3. Securing the Smart City IoT Infrastructure from QC and AI Threats

The timelines for AGI and QC align with the epoch when smart cities become the norm [29] [30]. Therefore, the existential threats from AGI and QC cannot be ignored in planning smart cities or any future digital infrastructure.

As illustrated in **Figure 8**, a smart city is made possible by the Internet of Things (IoT) [87], comprising a diverse range of computing devices that are inherently vulnerable [88]. Almost all cybersecurity currently relies on cryptography [89]. It is estimated that ~75 billion devices will be connected to the Internet by 2025 [90]. When QC arrives in the future, the security of the entire IT infrastructure will be threatened [24]. Post-quantum cryptography (PQC) is being aggressively pursued to defend the Internet against QC. NIST initiated a PQC standardization initiative in 2016-17, and after a rigorous multiyear vetting process, selected two out of 82 algorithms, CRYSTALS-Kyber and CRYSTALS-Dilithium. However, a Swedish group cracked CRYSTALS-Kyber [91] [92], and a group of French cryptographers recovered part of the secret key sufficient to produce

universal forgeries [93]. Moreover, a recent comprehensive survey confirms that the security of most PQC algorithms is unfortunately insufficient, rendering them vulnerable [94]. With no PQC algorithm proving robustness and resilience, NIST's standardization process is seriously jeopardized. Consequently, the original NIST projected timeline for PQC implementation is also seriously disrupted. As illustrated in **Figure 9**, updated from a recent report [38], all the

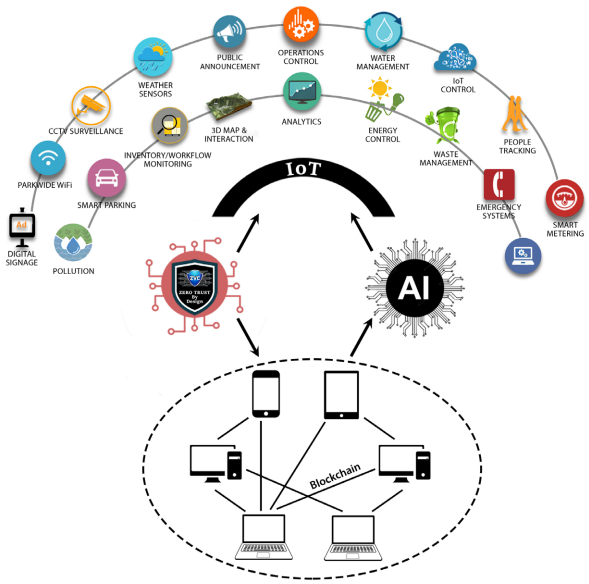

Figure 8. The concept of CASI enabled secure & controllable IoT infrastructure.

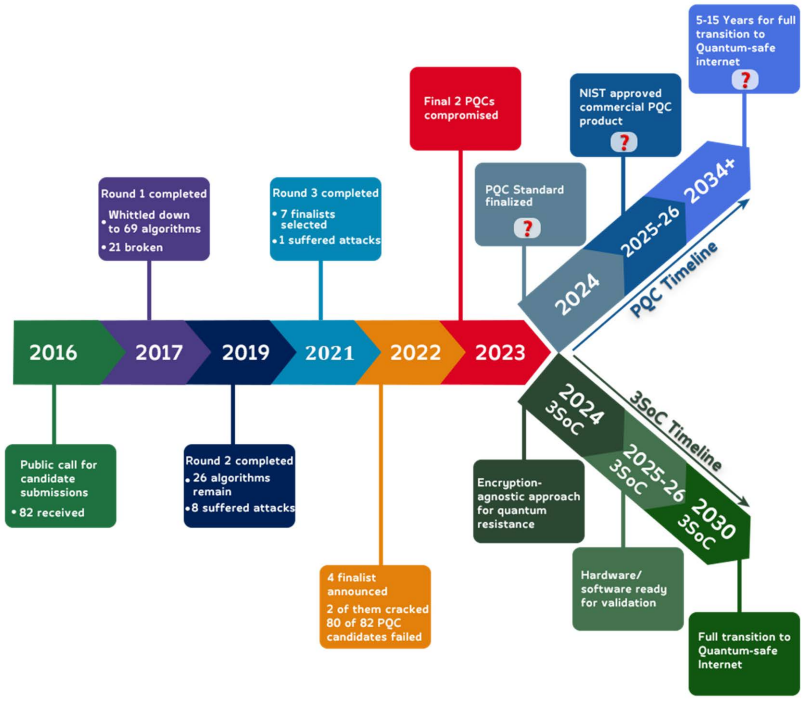

Figure 9. Timelineto build a defense against QC's threat to the Internet. Adapted from [39].

milestones beyond 2023 pose a serious question about the future of PQC in securing the Internet from QC's existential threat. Moreover, even if any PQCs reach production-grade standardization, most PQC algorithms are too complex to deploy efficiently in most resource-constrained IoT devices [95]. However, in peer-reviewed literature, empirical evidence exists [36] [37] [38] [39] that a resource-efficient, low-cost, encryption-agnostic approach based on banning all third-party permissions can be developed to secure network devices against Q-Day threats [38].

6. The Prospects and Limitations

The perspective presented in this paper is based on peer-reviewed empirical evidence that supports the proposed hypothesis on a secure and safe transition of AI and the Internet to the future human-friendly AGI and QC. Although the hypothesis has far-reaching implications for understanding computer designs, cybersecurity, and machine learning models' resilience, it remains under investigation. This paper is no more than hypothesis-generating research intended to build and formula tea hypothesis that researchers worldwide can design and investigate experiments to test and prove or disprove in the near future. Until such studies are conducted, great care should be taken to extrapolate the findings of this report to real-world settings. Such studies begin with defining what exactly is meant by quelling existential threats from AGI/QC. What will be the process? What protocols will be designed to implement the process? What KPIs will be appropriate to evaluate and control the protocols? Those and many other questions come to mind when planning the future of our AI/QC-powered digital infrastructure. The journey to answers to those questions largely depends on the following [38]:

i) The evolution of the business model *for delivering AI/AGI and QC services to the end users.*

ii) The safeguards that secure the Internet from *the AGI/QC threats relevant to that business model.*

Each of these elements is discussed in detail herein.

6.1. AI-as-a-Service Business Model

As ML models and neural networks constantly evolve, improve, and learn from new data, they are becoming more complex and resource-intensive. These advancements have resulted in AI services being offered in a pay-as-you-go cloud-based service model [96]. Such an AI-as-a-service business model is cost-effective and is rapidly establishing itself as a popular business model for providing users with pre-trained and optimized ML models [97].

6.2. Quantum-as-a-Service Model

The cost of building a QC is astronomically high, and it is impossible for most end users to buy or build one for their exclusive use. Therefore Quantum-

as-a-Service (QaaS) business model is the only choice for commercializing QC services. Several QC serviceshave already launched their QaaS product, offering their QC services to specialized groups building QC-based solutions [38].

6.3. Safeguards

Compared to traditional business practices, the "as-a-service" business model generally carries higher commercial viability because of cost savings. However, in the case of AI and QC, the respective business models provide an additional advantage of the ease of implementing CASI style safe, secure, and ethical framework, and makes regulatory control much more effective and enforceable because it is logistically easier to regulate business than the population at large [38].

Both the computing rules that this report challenges have deep roots in our practice of computing since the inception of the field of computer science and cannot be deracinated overnight. However, necessity is not only the mother of invention; it also mothers change. If the necessity is to save ourselves from extinction and change is the only choice left for humanity to survive, change will be inevitable. Time will tell if that change comes. However, until the hypothesis is proven with tangible evidence from multiple AI and QC labs, the hypothesis remains a concept.

Nevertheless, despite its limitation as hypothesis-generating research, this paper adds compelling evidence that controlling AI/AGI and QC is theoretically possible by using a new approach of encryption-agnostic decentralized governance to secure and control data for keeping it in compliance with ethical norms. The concept does demonstrate reasonable prospects of a credible path to be pursued by AI/QC researchers in their efforts to build solutions that mitigate any existential risk that AGI/QC may pose in the future. It also holds out new hopes of a smooth passage to the technological singularity when it arrives, without questioning the debate on whether singularity will arrive or will not [98].

7. Discussion & Conclusion

The principal objective of this research was to identify the most serious pain points posed by AI/AGI and QC technologies that are currently perceived as existential risks to humanity by many experts [23] [35] [36], and conduct a thorough literature review to generate a clearly articulated hypothesis that provides a credible path to mitigating the involved risks. The hypothesis thus formulated reads as follows:

"Safe, secure, ethical, and controllable AGI/QC is possible by conquering the two unassailable rules of computability with Collective Artificial Super Intelligence (CASI)."

The empirical evidence in peer-reviewed literature provided enough basis to support the above hypothesis and afford sufficient motivation to AI and QC researchers to undertake further research to test and prove the hypothesis. This

work introduces new ideas, new thinking, and a new understanding of the paradoxical computability concepts of permission-based Turing machines that have existed since the birth of modern computers. The new perspective on those age-old concepts can be helpful to researchers, thinkers, AI developers, regulators, and practitioners working to secure the Internet generally and the brand-new fields of AI and QC specifically.

The Pause-AI call signatories believed the 6-month moratorium would give AI companies and regulators time to formulate safeguards to protect society from potential risks of the technology [12]. Conversely, it is the dynamic input from research labs that decides the path that regulators take. Microsoft co-founder Bill Gates told Reuters the proposed pause will not "*solve the challenges* [12]". The need of the hour is "*acceleration, not a pause* [99]". Towing a similar but more diplomatic line, the Google CEO characterized the call as "a conversation starter" backed by good spirit [100]. With tech luminaries on both sides not disputing the potentially catastrophic dangers of uncontrolled AI/AGI and QC, the Pause-AI call is indeed a conversation starter of epoch magnitude. This paper is testimony to that epoch for providing an optimistic research direction to the AI stakeholders on either side of the debate.

Besides the wildly raging AI controversy, the PQC standardization process initiated by NIST in 2016-17 as a defense against QC is also going through tough times, with 90% of PQC algorithms failing in the fourth round [89], and recently, the final two PQC algorithms, viz. CRYSTALS (Kyber & Dilithium [91] [92]) also reported to be compromised. With these setbacks, both AGI and QC remain defenseless. As researchers continue to find solutions to these intractable problems, the CASI hypothesis merits a pursuit as a more sustainable alternative to PQC. The hypothesis can be further broken down into two key research questions as follows:

i) Will the integration of blockchain render AI controllable?

ii) Can ZVC (Zero Vulnerability Computing) provide autonomous and seamless security to AI and Quantum Computing?

Indeed, as we speak, CASI and these research questions are actively pursued by a consortium of European researchers. In these challenging times, the alternate research direction that the CASI hypothesis proposed and supported with empirical evidence in this paper may go a long way in planning our defenses against the impending dangers to our digital infrastructures from bad actors' future abuse of AGI and QC.

CRediT Authors' Contribution Statement

FR is the sole contributor to this study's hypothesis generation, support, and testing. The author has read and agreed to the published version of the manuscript.

Acknowledgements

The author is grateful to Dr Brecht Vermeulen and Professor Peter Van Daele

(IMEC-Ghent University, IDLabiGent Tower-Department of Information Technology, Technologiepark-Zwijnaarde 126, B-9052 Ghent, Belgium) for support in the initial hypothesis building research, and to Mr Tejas Bhagat and Ms Saidya Khan for their help in preparing this manuscript. The author is also grateful to DrKotzanikolaou Panayiotis of the University of Piraeus and DrKostas Kolomvatsos of the University of Thessaly for being trusted partners in several consortia and initiatives involved in the development of the Zero Vulnerability Computing (ZVC) concept.

Conflicts of Interest

The author declares that he have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Peters, M.A., *et al.* (2023) AI and the Future of Humanity: ChatGPT-4, Philosophy and Education—Critical Responses. *Educational Philosophy and Theory*, 1-35. https://doi.org/10.1080/00131857.2023.2213437

[2] Ford, P. (2023) The Quantum Cybersecurity Threat May Arrive Sooner than You Think. *Computer*, **56**, 134-136. https://doi.org/10.1109/MC.2022.3227657

[3] Kline, K., Salvo, M. and Johnson, D. (2019) How Artificial Intelligence and Quantum Computing Are Evolving Cyber Warfare. Cyber Intelligence Initiative, the Institute of World Politics, Washington, DC.

[4] Chen, L., *et al.* (2016) Report on Post-Quantum Cryptography. Department of Commerce and National Institute of Standards and Technology, Gaithersburg, MD. https://doi.org/10.6028/NIST.IR.8105

[5] Moret-Bonillo, V. (2015) Can Artificial Intelligence Benefit from Quantum Computing? *Progress in Artificial Intelligence*, **3**, 89-105. https://doi.org/10.1007/s13748-014-0059-0

[6] Acampora, G. (2019) Quantum Machine Intelligence: Launching the First Journal in the Area of Quantum Artificial Intelligence. *Quantum Machine Intelligence*, **1**, 1-3. https://doi.org/10.1007/s42484-019-00006-5

[7] Marr, B. (2023) A Short History of ChatGPT: How We Got Where We Are Today. https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today/

[8] Buriak, J.M., *et al.* (2023) Best Practices for Using AI When Writing Scientific Manuscripts: Caution, Care, and Consideration: Creative Science Depends on It. *ACS Nano*, **17**, 4091-4093. https://doi.org/10.1021/acsnano.3c01544

[9] Yu, H. (2023) Reflection on Whether Chat GPT Should Be Banned by Academia from the Perspective of Education and Teaching. *Frontiers in Psychology*, **14**, Article 1181712. https://doi.org/10.3389/fpsyg.2023.1181712

[10] Bubeck, S., *et al.* (2023) Sparks of Artificial General Intelligence: Early Experiments with GPT-4. arXiv Preprint arXiv:2303.12712

[11] Blake, A. (2023) GPT-5 Could Change the World in One Incredible Way. https://www.digitaltrends.com/computing/gpt-5-artificial-general-intelligence/

[12] Clarke, L. (2023) Call for AI Pause Highlights Potential Dangers. *Science*, **380**,

120-121. https://doi.org/10.1126/science.adi2240

[13] Future of Life Institute (2023) Pause Giant AI Experiments: An Open Letter. https://futureoflife.org/open-letter/pause-giant-ai-experiments/

[14] Kevin, R. (2023) A.I. Poses 'Risk of Extinction,' Industry Leaders Warn. https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html

[15] Taylor, H. (2023) Ministers Not Doing Enough to Control Ai, Says UK Professor. https://www.theguardian.com/technology/2023/may/13/ministers-not-doing-enough-to-control-ai-says-uk-professor

[16] Tredinnick, L. and Laybats, C. (2023) The Dangers of Generative Artificial Intelligence. *Business Information Review*, **40**, 46-48. https://doi.org/10.1177/02663821231183756

[17] Ambartsoumean, V.M. and Yampolskiy, R.V. (2023) AI Risk Skepticism, a Comprehensive Survey. arXiv Preprint arXiv:2303.03885

[18] Samuel, J. (2023) Response to the March 2023 'Pause Giant AI Experiments: An Open Letter' by Yoshua Bengio, Signed by Stuart Russell, Elon Musk, Steve Wozniak, Yuval Noah Harari and Others… *SSRN Electronic Journal*. https://ssrn.com/abstract=4412516

[19] Richter, F. (2023) Will AI Go Rogue? https://www.statista.com/chart/29514/fear-of-artificial-intelligence-going-rogue/

[20] O'Carroll, L. (2023) EU Moves Closer to Passing One of World's First Laws Governing AI. https://www.theguardian.com/technology/2023/jun/14/eu-moves-closer-to-passing-one-of-worlds-first-laws-governing-ai

[21] Kuusi, O. and Heinonen, S. (2022) Scenarios from Artificial Narrow Intelligence to Artificial General Intelligence—Reviewing the Results of the International Work/Technology 2050 Study. *World Futures Review*, **14**, 65-79. https://doi.org/10.1177/19467567221101637

[22] Sasi, P., *et al.* (2023) Quantum Computing and the Qubit: The Future of Artificial Intelligence. In: Tyagi, A., Ed., *Handbook of Research on Quantum Computing for Smart Environments*, IGI Global, Hershey, 231-244. https://doi.org/10.4018/978-1-6684-6697-1.ch013

[23] Mallow, G.M., *et al.* (2022) Quantum Computing: The Future of Big Data and Artificial Intelligence in Spine. *Spine Surgery and Related Research*, **6**, 93-98. https://doi.org/10.22603/ssrr.2021-0251

[24] Schiffer, B.F. (2022) Quantum Computers as an Amplifier for Existential Risk. arXiv Preprint arXiv:2205.02761

[25] Grimes, R.A. (2019) Cryptography Apocalypse: Preparing for the Day When Quantum Computing Breaks Today's Crypto. John Wiley & Sons, Hoboken. https://doi.org/10.1002/9781119618232

[26] Fernandez-Carames, T.M. and Fraga-Lamas, P. (2020) Towards Post-Quantum Blockchain: A Review on Blockchain Cryptography Resistant to Quantum Computing Attacks. *IEEE Access*, **8**, 21091-21116. https://doi.org/10.1109/ACCESS.2020.2968985

[27] Gomes, L. (2018) Quantum Computing: Both Here and Not Here. *IEEE Spectrum*, **55**, 42-47. https://doi.org/10.1109/MSPEC.2018.8322045

[28] Schmierer, R. (2022) Post Quantum Computing Survey Results—Are You Ready? https://itchronicles.com/technology/post-quantum-computing-survey-results/

[29] Huttner, B. and Kalsi, M. (2022) Countdown to Y2Q: Working Group, Quantum-

Safe Security.
https://cloudsecurityalliance.org/research/working-groups/quantum-safe-security/

[30] Pozoukidou, G. and Angelidou, M. (2022) Urban Planning in the 15-Minute City: Revisited under Sustainable and Smart City Developments until 2030. *Smart Cities*, **5**, 1356-1375. https://doi.org/10.3390/smartcities5040069

[31] UNESCO and NETEXPLO (2019) Smart Cities: Shaping the Society of 2030. United Nations Educational, Scientific and Cultural Organization (UNESCO), Paris.

[32] Gabor, T., *et al*. (2020) The Holy Grail of Quantum Artificial Intelligence: Major Challenges in Accelerating the Machine Learning Pipeline. *Proceedings of the IEEE/ACM* 42*nd International Conference on Software Engineering Workshops*, Seoul, 27 June-19 July 2020, 456-461. https://doi.org/10.1145/3387940.3391469

[33] Andreasson, A., *et al*. (2020) A Census of Swedish Government Administrative Authority Employee Communications on Cybersecurity during the COVID-19 Pandemic. 2020 *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (*ASONAM*), The Hague, 7-10 December 2020, 727-733. https://doi.org/10.1109/ASONAM49781.2020.9381324

[34] Pal, K. (2023) Can Quantum Computing Impact the Applications of Artificial Intelligence.
https://www.techopedia.com/can-quantum-computing-impact-the-applications-of-artificial-intelligence

[35] Majot, A. and Yampolskiy, R. (2015) Global Catastrophic Risk and Security Implications of Quantum Computers. *Futures*, **72**, 17-26.
https://doi.org/10.1016/j.futures.2015.02.006

[36] Raheman, F. (2022) The Future of Cybersecurity in the Age of Quantum Computing. *Future Internet*, **14**, Article 335. https://doi.org/10.3390/fi14110335

[37] Raheman, F., Bhagat, T., Vermeulen, B. and Van Daele, P. (2022) Will Zero Vulnerability Computing (ZVC) Ever Be Possible? Testing the Hypothesis. *Future Internet*, **14**, Article 238. https://doi.org/10.3390/fi14080238

[38] Raheman, F. (2022) The Q-Day Dilemma and the Quantum Supremacy/Advantage Conjecture. Research Square. https://doi.org/10.21203/rs.3.rs-2331935/v1

[39] Raheman, F. (2024) From Standard Policy-Based Zero Trust to Absolute Zero Trust (AZT): A Quantum Leap to Q-Day Security. *Journal of Computer and Communications*, **12**, 252-282. https://doi.org/10.4236/jcc.2024.123016

[40] Strachey, C. (1965) An Impossible Program. *The Computer Journal*, **7**, 313.
https://doi.org/10.1093/comjnl/7.4.313

[41] Alfonseca, M., *et al*. (2021) Superintelligence Cannot Be Contained: Lessons from Computability Theory. *Journal of Artificial Intelligence Research*, **70**, 65-76.
https://doi.org/10.1613/jair.1.12202

[42] Calude, C.S. and Dumitrescu, M. (2018) A Probabilistic Anytime Algorithm for the Halting Problem. *Computability*, **7**, 259-271. https://doi.org/10.3233/COM-170073

[43] Stoddart, B. (2019) The Halting Paradox. arXiv Preprint arXiv:1906.05340

[44] Hartwick, J. and Barki, H. (1994) Research Report—Hypothesis Testing and Hypothesis Generating Research: An Example from the User Participation Literature. *Information Systems Research*, **5**, 446-449. https://doi.org/10.1287/isre.5.4.446

[45] Biesecker, L.G. (2013) Hypothesis-Generating Research and Predictive Medicine. *Genome Research*, **23**, 1051-1053. https://doi.org/10.1101/gr.157826.113

[46] Saghiri, A.M., *et al*. (2022) A Survey of Artificial Intelligence Challenges: Analyzing the Definitions, Relationships, and Evolutions. *Applied Sciences*, **12**, Article 4054.

https://doi.org/10.3390/app12084054

[47] Russell, S. (2019) Human Compatible: Artificial Intelligence and the Problem of Control. Penguin, London.

[48] Yampolskiy, R.V. (2020) On Controllability of AI. arXiv:2008.04071

[49] Babcock, J., Kramar, J. and Yampolskiy, R.V. (2016) The AGI Containment Problem. *Proceedings of the* 9*th International Conference on Artificial General Intelligence*, New York, 16-19 July 2016, 53-63. https://doi.org/10.1007/978-3-319-41649-6_6

[50] Saltzer, J.H. and Schroeder, M.D. (1975) The Protection of Information in Computer Systems. *Proceedings of the IEEE*, **63**, 1278-1308. https://doi.org/10.1109/PROC.1975.9939

[51] Steinhardt, J., Koh, P.W. and Liang, P. (2017) Certified Defenses for Data Poisoning Attacks. *Proceedings of the* 31*st Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 3520-3532.

[52] Akhtar, N., *et al*. (2021) Advances in Adversarial Attacks and Defenses in Computer Vision: A Survey. *IEEE Access*, **9**, 155161-155196. https://doi.org/10.1109/ACCESS.2021.3127960

[53] Isaac, E.R.H.P. and Reno, J. (2023) AI Product Security: A Primer for Developers. arXiv Preprint ArXiv:2304.11087 https://arxiv.org/pdf/2304.11087.pdf

[54] Tatam, M., *et al*. (2021) A Review of Threat Modeling Approaches for APT-Style Attacks. *Heliyon*, **7**, E05969. https://doi.org/10.1016/j.heliyon.2021.e05969

[55] Xiong, W. and Lagerström, R. (2019) Threat Modeling—A Systematic Literature Review. *Computers & Security*, **84**, 53-69. https://doi.org/10.1016/j.cose.2019.03.010

[56] Shostack, A. (2022) Threat Modeling: Designing for Security. Wiley, Hoboken.

[57] Mauri, L. and Damiani, E. (2022) Modeling Threats to AI-ML Systems Using STRIDE. *Sensors*, **22**, Article 6662. https://doi.org/10.3390/s22176662

[58] Mauri, L. and Damiani, E. (2021) STRIDE-AI: An Approach to Identifying Vulnerabilities of Machine Learning Assets. 2021 *IEEE International Conference On Cyber Security and Resilience* (*CSR*), Rhodes, 26-28 July 2021, 147-154. https://doi.org/10.1109/CSR51186.2021.9527917

[59] European Union Agency for Cybersecurity, Malatras, A. and Dede, G. (2023) AI Cybersecurity Challenges: Threat Landscape for Artificial Intelligence. https://op.europa.eu/en/publication-detail/-/publication/e52bf2d7-4017-11eb-b27b-01aa75ed71a1/language-en

[60] Cook, B., Podelski, A. and Rybalchenko, A. (2011) Proving Program Termination. *Communications of the ACM*, **54**, 88-98. https://doi.org/10.1145/1941487.1941509

[61] Lucas, S. (2021) The Origins of the Halting Problem. *Journal of Logical and Algebraic Methods in Programming*, **121**, Article 100687. https://doi.org/10.1016/j.jlamp.2021.100687

[62] Dietrich, E. and Fields, C. (2020) Equivalence of the Frame and Halting Problems. *Algorithms*, **13**, Article 175. https://doi.org/10.3390/a13070175

[63] Rybalov, A. (2007) On the Strongly Generic Undecidability of the Halting Problem. *Theoretical Computer Science*, **377**, 268-270. https://doi.org/10.1016/j.tcs.2007.02.010

[64] Gams, M. (2013) Alan Turing, Turing Machines and Stronger. *Informatica*, **37**, 9-14.

[65] Hyun, W.-S. (2012) Turing's Cognitive Science: A Metamathematical Essay for His Centennial. *Korean Journal of Cognitive Science*, **23**, 367-388. https://doi.org/10.19066/cogsci.2012.23.3.004

[66] Fairfield, J.A.T. (2019) The Human Element: The Under-Theorized and Underutilized Component Vital to Fostering Blockchain Development. *Cleveland State Law Review*, **67**, 33-41.

[67] Cardona, R., Miranda, E. and Peralta-Salas, D. (2021) Looking at Euler Flows through a Contact Mirror: Universality and Undecidability. arXiv Preprint arXiv:2107.09471

[68] Brennan, L. (2023) AI Ethical Compliance Is Undecidable. *Hastings Science and Technology Law Journal*, **14**, 311-338.

[69] Floridi, L. and Cowls, J. (2022) A Unified Framework of Five Principles for AI in Society. In: Carta, S., Ed., *Machine Learning and the City: Applications in Architecture and Urban Design*, John Wiley & Sons Ltd., Hoboken, 535-545. https://doi.org/10.1002/9781119815075.ch45

[70] Anderson, S.L. (2008) Asimov's Three Laws of Robotics and Machine Metaethics. *AI & Society*, **22**, 477-493. https://doi.org/10.1007/s00146-007-0094-5

[71] Zenil, H. (2013) A Behavioural Foundation for Natural Computing and a Programmability Test. In: Dodig-Crnkovic, G. and Giovagnoli, R., Eds., *Computing Nature: Turing Centenary Perspective*, Springer, Berlin, 87-113. https://doi.org/10.1007/978-3-642-37225-4_5

[72] Amos, Z. (2022) Data Poisoning: Is There a Solution? https://www.unite.ai/data-poisoning-is-there-a-solution/

[73] European Commission (2023) "Seal of Excellence" Awarded to ZVC in a Horizon Europe EIC Accelerator Grant Program. https://zvchub.com/#seal

[74] Darwish, D. (2023) Blockchain and Artificial Intelligence for Business Transformation toward Sustainability. In: Namasudra, S. and Akkaya, K., Eds., *Blockchain and Its Applications in Industry* 4.0, Springer, Singapore, 211-255. https://doi.org/10.1007/978-981-19-8730-4_8

[75] Badruddoja, S., *et al.* (2022) Making Smart Contracts Predict and Scale. 2022 *Fourth International Conference on Blockchain Computing and Applications* (*BCCA*), San Antonio, 5-7 September 2022, 127-134. https://doi.org/10.1109/BCCA55292.2022.9922480

[76] Leontaris, L., *et al.* (2023) A Blockchain-Enabled Deep Residual Architecture for Accountable, *In-Situ* Quality Control in Industry 4.0 with Minimal Latency. *Computers in Industry*, **149**, Article 103919. https://doi.org/10.1016/j.compind.2023.103919

[77] Du, Y., Wang, Z. and Leung, V.C.M. (2021) Blockchain-Enabled Edge Intelligence for IoT: Background, Emerging Trends and Open Issues. *Future Internet*, **13**, Article 48. https://doi.org/10.3390/fi13020048

[78] Zhong, L., Qi, C. and Gao, Y. (2022) Blockchain-Enabled Automatic Learning Method for Digital Gaming Systems Based on Big Data. *International Journal of Gaming and Computer-Mediated Simulations* (*IJGCMS*), **14**, 1-22. https://doi.org/10.4018/ijgcms.315634

[79] Kaushik, K. (2022) Blockchain Enabled Artificial Intelligence for Cybersecurity Systems. In: Ouaissa, M., Boulouard, Z., Ouaissa, M., Khan, I.U. and Kaosar, M., Eds., *Big Data Analytics and Computational Intelligence for Cybersecurity*, Springer International Publishing, Cham, 165-179. https://doi.org/10.1007/978-3-031-05752-6_11

[80] Shinde, R., *et al.* (2022) Securing AI-Based Healthcare Systems Using Blockchain Technology: A State-of-the-Art Systematic Literature Review and Future Research

Directions. *Transactions on Emerging Telecommunications Technologies*, **35**, e4884. https://doi.org/10.1002/ett.4884

[81] Shen, M. *et al.* (2023) Blockchains for Artificial Intelligence of Things: A Comprehensive Survey. *IEEE Internet of Things Journal*, **10**, 14483-14506. https://doi.org/10.1109/JIOT.2023.3268705

[82] Junis, F., *et al.* (2019) A Revisit on Blockchain-Based Smart Contract Technology. arXiv Preprint arXiv:1907.09199

[83] Eberhardt, J. and Tai, S. (2018) Zokrates-Scalable Privacy-Preserving Off-Chain Computations. 2018 *IEEE International Conference on Internet of Things* (*IThings*) *and IEEE Green Computing and Communications* (*GreenCom*) *and IEEE Cyber, Physical and Social Computing* (*CPSCom*) *and IEEE Smart Data* (*SmartData*), Halifax, 30 July-3 August 2018, 1084-1091. https://doi.org/10.1109/Cybermatics_2018.2018.00199

[84] Jansen, M., *et al.* (2020) Do Smart Contract Languages Need to Be Turing Complete? *Proceedings of the* 1*st International Congress on Blockchain and Applications* 2019, Ávila, 26-28 June 2019, 19-26. https://doi.org/10.1007/978-3-030-23813-1_3

[85] Hu, B., *et al.* (2021) A Comprehensive Survey on Smart Contract Construction and Execution: Paradigms, Tools, and Systems. *Patterns*, **2**, Article 100179. https://doi.org/10.1016/j.patter.2020.100179

[86] Mintlayer (2023) Why DeFi's Future Is with Non-TuringComplete Smart Contracts. https://www.mintlayer.org/news/2020-11-05-why-defis-future-is-with-non-turing-complete-smart-contracts/

[87] Jin, J., *et al.* (2014) An Information Framework for Creating a Smart City through Internet of Things. *IEEE Internet of Things Journal*, **1**, 112-121. https://doi.org/10.1109/JIOT.2013.2296516

[88] DrFazal (2023) Why Computers Are Inherently Vulnerable? https://drfazal.medium.com/why-computers-are-inherently-vulnerable-fd7a34afaec6

[89] Liu, X., *et al.* (2018) Biometrics-Based RSA Cryptosystem for Securing Real-Time Communication. *Sustainability*, **10**, Article 3588. https://doi.org/10.3390/su10103588

[90] Aljabri, M.G. (2023) Blockchain Technology. In: Srivastava, D., Sharma, N., *et al*, Eds., *Intelligent Internet of Things for Smart Healthcare Systems*, CRC Press, Boca Raton, 165. https://doi.org/10.1201/9781003326182-11

[91] Townsend, K. (2023) AI Helps Crack NIST-Recommended Post-Quantum Encryption Algorithm. https://www.securityweek.com/ai-helps-crack-a-nist-recommended-post-quantum-encryption-algorithm/

[92] Ji, Y. and Dubrova, E. (2023) A Side-Channel Attack on a Masked Hardware Implementation of CRYSTALS-Kyber. *Proceedings of the* 2023 *Workshop on Attacks and Solutions in Hardware Security*, Copenhagen, 30 November 2023, 27-37. https://eprint.iacr.org/2023/1084 https://doi.org/10.1145/3605769.3623992

[93] Berzati, A., *et al.* (2023) A Practical Template Attack on CRYSTALS-Dilithium. Cryptology ePrint Archive.

[94] Canto, A.C., *et al.* (2023) Algorithmic Security Is Insufficient: A Comprehensive Survey on Implementation Attacks Haunting Post-Quantum Security. arXiv Preprint arXiv:2305.13544

[95] Hadayeghparast, S., Bayat-Sarmadi, S. and Ebrahimi, S. (2022) High-Speed Post-Quantum Cryptoprocessor Based on RISC-V Architecture for IoT. *IEEE Internet of Things Journal*, **9**, 15839-15846. https://doi.org/10.1109/JIOT.2022.3152850

[96] Lewicki, K.L., *et al.* (2023) Out of Context: Investigating the Bias and Fairness Concerns of "Artificial Intelligence as a Service". *Proceedings of the* 2023 *CHI Conference on Human Factors in Computing Systems*, Hamburg, 23-28 April 2023, 1-17. https://doi.org/10.1145/3544548.3581463

[97] Chethana, C., Shaik, M. and Pareek, P. (2023) Artificial Intelligence Applications for Process Optimization in Small Software Firms. *SSRN Electronic Journal*.

[98] Tariq, S., *et al.* (2023) Is the 'Technological Singularity Scenario' Possible: Can AI Parallel and Surpass All Human Mental Capabilities? *World Futures*, **79**, 200-266. https://doi.org/10.1080/02604027.2022.2050879

[99] Mathews, J. (2023) Microsoft's Chief Scientific Officer, One of the World's Leading A.I. Experts, Doesn't Think a 6 Month Pause Will Fix A.I.—But Has Some Ideas of How to Safeguard It. https://fortune.com/2023/04/30/microsoft-eric-horvitz-ai-research-predictions/

[100] Mollman, S. (2023) Google CEO Won't Commit to Pausing A.I. Development after Experts Warn about 'Profound Risks to Society'. https://fortune.com/2023/03/31/google-ceo-sundar-pichai-artificial-intelligence-open-letter-response/