Intelligent Virtual Assistants (IVAs): Trust and Zero Trust

Allison Wylde

Cardiff University, Cardiff, Wales, CF10 3EU, UK

wyldea@cardiff.ac.uk

Abstract. Intelligent virtual assistants (IVAs) for cyber security appear to offer promising solutions to tackle the problem of gaps in the future cyber security workforce. However, this paper argues that a problem emerges as artificial intelligence (AI) partners take on their roles. In AI - implicit trust is the norm, yet in cyber security, zero trust protocols are now mandated. The contribution of this conceptual paper is firstly to present an argument for the deployment of zero trust protocols to effectively manage our future AI partners, and secondly, to set out the first steps in a process to assess the operationalization of zero trust. By leveraging well-established theory on trust from organization and conflict management studies, zero trust can be evaluated. The zero trust assessment involves determining: propensity to trust; experience of trust; a trust assessment based on ability, benevolence, and integrity; followed by a study of; acceptance of vulnerability and of risk taking. Implications for practitioners, policymakers and academics include an argument that the deployment, assessment, and management of a zero trust posture will promote explainable, trustworthy, and secure AI. Further studies are called for.

Keywords: Artificial Intelligence, AI, Trust, Zero Trust, Cyber Security.

1 Introduction

Worldwide the impact of cyber criminals and cyber crime is increasing in frequency scale and impact. At the same time estimates of gaps in the cyber security workforce in 2022, are estimated to be more than 3.4 million cyber security practitioners [1]. The focus of this paper is to explore aspects of trust in the operations of AI trained intelligent virtual assistants (IVAs) as they are deployed for cyber security.

Although much is known about AI, IVAs, and trust, to the best of the author's knowledge little is published regarding how to assess trust in a context of cyber security operations (involving IVAs) that require mandated zero trust protocols. What follows in this paper is not a systematic survey of state of the art, rather a review of key aspects in the context of trying to solve the problem.

The contribution of the paper is at the intersection between AI, trust, and cyber security. Through presenting an argument for the adoption of zero trust in the operations of AI and IVAs deployed for cyber security this paper sets out a fresh approach and a new process to allow assessment of the operationalization of zero trust.

The paper is structured as follows. To achieve the goal of furthering our understanding, this paper proceeds by first, in section 1, the introduction, presenting the reader with the central argument that zero trust is an important and overlooked protocol essential for cyber security. In section 2, a discussion on AI with a particular focus on IVAs is presented. The third section trust theory from management and organization studies is reviewed with a focus on the key elements involved in trust building developed further to help to understanding. Section 4 discusses zero trust, in the context of standards and guidelines, and zero trust protocols for AI, concerning IVAs deployed in operation. This work is part of a larger study which also examines aspects of workforce acceptance, however, due to space limitations, this is not covered in this paper. In the final section, 6, the conclusion, important implications for practitioners, policymakers and academics are presented.

2 AI and Intelligent Virtual Assistants (IVAs)

AI is not new, in fact in banking and in healthcare, the next section summarizes AI and IVAs along with some unforeseen risks and challengers with their adoption.

AI has been in use since the 1950s with systems able to make inferences and later in the 1960s and 70s, to conduct searches for, and the retrieval of medical literature [2]. Later developments helped improve diagnostic accuracy and procedural accuracy that resulted in improved patient outcomes [2]. IVAs were developed from chat bots, which themselves were originally developed to provide on-line conversation through text or speech to simulate a human [3]. IVAs are now more sophisticated using natural language processing (NLP) to match users' words, whether as text or spoken and to process images [4]. Many IVAs use AI to learn through searches, lookups, backend databases, rules, and reference engines that provide answers, which may not or not always be reliable [5].

The development of IVAs has continued with more recent examples, such as developed by Apple, Siri in an iPhone, Microsoft Amazon, Alexa and Echo device, and Google Assistant on Google enables Android devices.

Activating the IVA's listening, recording (voice or image) and processing functions requires a call or wake signal such as 'Hey Siri', 'OK Google' or 'Alexa'. Once activated the IVA carries out tasks, often involving communicating with multiple devices, such as smart speakers or headphones, or smart home fridges, thermostats, lights, and security systems [5]. The IVAs are also linked with third-party vendors, allowing bank balances to be checked and items, such as, pizzas, video streaming or ride hires to be purchased [5]. In October 2022, a humanoid android device named Ai-Da attended a high-level UK parliament committee hearing and provided evidence through using a language mode [6].

Not surprisingly several risks have emerged, notably [5] who explore the ecosystem of IVAs and highlight risks such as eavesdropping, voice recording and hacking, the authors call for the need to understand growing security and privacy threats. Key issues

privacy issues are described by [7] as the 'end of privacy'. Security threats include malicious commands which could result in unwanted ordering, or attacks on property through opening doors and allowing thefts. Unintentional ordering of products has occurred, in one example a six-year-old in Dallas, Texas, ordered a doll's house (Sparkle Mansion) and cookies (snacks), much to her parents' surprise when they arrived [5]. Other instances include recordings of private conversations being used by Apple and Google for product training purposes [8].

Several cyber security challenges have been identified so far in the previous incidents, the main problems are arguably linked to the current procedures based on presumptive trust in the systems and networks. These incidents have resulted in unplanned purchases, inconvenience and at worst privacy violations [8]. However, in cases where critical processes or high-risk AI systems or environments are involved, such as healthcare, financial services or in nuclear power generation, events that are not planned or unexpected, such as cyber security breaches or actions, could prove fatal, economically damaging and/ or catastrophic.

The impacts of high-risk incidents could extend beyond a single individual, organization or country. In a nuclear event, the outcome could extend across several nations or indeed in the worst case, world-wide. It is thus essential to understand the basic trust functions and assumptions and to create further awareness of potential cyber security threats as well provide approaches that allow for monitoring and prevention.

For this paper, the focus is on understanding the nature of and the role of trust- discussed next.

3 Trust

The key assumptions and elements in the construct of trust are well-researched. Trust has also been characterized as operating in and at different levels and between different respondents [9]. As the paper is not a systematic survey, the focus here is trying to arrive at a framework for analysis that is applicable to the context of virtual assistants.

Two key approaches from trust research are drawn on to create a framework through which trust can be examined. First the integrative trust formation model [10] from organization and management studies is examined followed by a consideration of important components from conflict resolution studies [11]. What is presented is not a complete state of the art, but rather a review of specific material, in the context of trying to create a frame to understand the puzzle at the heart of this paper.

Trust is widely viewed as multi-faceted and based on several elements. These include the presence of positive expectations of a trustee's trustworthiness and an assessment of trustworthiness [10]. Next, trust is assessed based on three components, ability (does an individual have the ability necessary to perform a particular action?), benevolence (does the individual act in good conscience?) and lastly, integrity (does the individual act with integrity?) [10]. This assessment is moderated by a trustor's; propensity to trust; willingness to accept vulnerability and to take risk, in the relationship (on the part of the trustor) [10]. From conflict resolution trust development is viewed as based on the foundations of a trustor's willingness to act and their ability to trust as moderated by trust experiences [11].

Table 1. A throughput model of trust building, expanding [10] and [11].

Trustor's antecedents	Trustor's assessment	Trustor's actions
Propensity	Propensity	Propensity
Ability to trust	Ability	Acceptance of vulnerability
Trust, prior experience	Benevolence	Assessment of risk
	Integrity (ABI)	Risk taking behavior

The review above gives rise to a standard definition of trust as based on a trustor's positive expectations and willingness to be vulnerable to the actions of the trustee, with an expectation that the trustee will undertake an action important to the trustor, irrespective of control or monitoring by the trustor [10]. In this paper trust is conceptualized based on the definition as presented. Importantly, trust is linked to and tied into an individual trustor's propensity to trust [10], personality, belief systems and trust experiences [11], set out in Table 1, above.

Trust has also been studied at different levels and different referents, in teams, organizations and institutions [9]. Trust has also been examined in non-person based relations, for example, trust in a policy [10] or a technology [12]. What is important in the context of this paper are ideas that trust exists in entities outside of, and indeed, beyond human relations [12]. Considered next are a definition for zero trust, and zero trust operations to help understanding.

4 Zero Trust

The cyber security approach zero trust was first proposed in 2010 [13]. Zero trust counters an over-dependence on, and the presence of presumptive trust and trusted systems [13],[14]. Key views of zero trust are discussed next.

In the context of cyber security, trust is viewed as a vulnerability [13]. In this paper zero trust is based no presumptive trust, rather a risk-based approach to granting trust [13]. Zero trust relies on continuous monitoring and verification [13]. The zero trust approach deals with limitations in traditional fixed boundary or perimeter-based trust approaches. Trust cannot be granted based on location, in fact in modern network, cloud based, and internet of things (IoT) organizational boundaries no longer exist [13],[14].

Recent policy and guidelines from international bodies and governments promote and, in in the US, zero trust is mandated [15]. Guidance on zero trust implementation involves the verification of identity, individual, device, process, and or service, such as IoT [14]. As well as an assessment of context and state [13].

The NIST definition of zero trust involves: minimizing uncertainty and enforcing decisions based on least privilege access peer-request-access in information systems in the face of a network, which is viewed as compromised [16]. A zero trust architecture therefore comprises an enterprise's cyber security plan (zero trust based), component relationships, workflow planning, and access policies [16]. In sum, a NIST zero trust enterprise is the sum of the physical and virtual network infrastructure and zero trust policies [16]. NIST define the terms, user, subject, and resource, as entities that may

request information from resources (assets, applications, workflows, network accounts, services, and devices) that may substitute as data [16]. Interestingly, the term user is reserved for humans, while subject is the standard term for all other entities [16]. Zero trust minimizes access to identified subjects and assets requiring access, based on an authentic subject/ user and valid request - while continuously authenticating and authorizing each request [16]. The process, referred to as policy decision and policy enforcement policy (PDP/ PEP) judgements, may be managed by trust algorithms [16]. NIST also adds that zero trust was in operation long before it was named zero trust [16].

The UK NCSC's ten principles for zero trust, include: knowledge of architecture; the creation of a single strong identity; strong device identity; authentication of everything; no trust in any network; and the selection of services designed for zero trust [14]. The NCSC's approach involves policy and continuous, authorized decision-making to help in the practical implementation of zero trust [14]. Prominent approaches and protocols in zero trust application such as zero knowledge and garbled circuits [17], are not discussed here due to limitations of scope.

Operationalizing zero trust therefore relies on continuous decision-making and monitoring to ensure that confidential and sensitive information is not discoverable. Next, the application of trust and zero trust in the context of AI and AVIs is discussed.

5 IVA's: Trust, Zero Trust

As highlighted above, AVI's are in operation now, and use is set to grow in the domain of cyber security. What this paper suggests is that in light of the challenges highlighted earlier, it is necessary to address the pressing issue of trust. Trust in this domain is explored next. As a starting point one emerging policy is considered, the October 2022, US Government's AI Bill of Rights [18][19]. This is followed by a consideration of the expanded trust models presented above, with the context of zero trust included.

At the time of writing, the 2022, US Government's AI Bill of Rights [18],[19] has just been released, setting out five principles (underpinned by trust) to ensure that AI is trusted and trustworthy and protects the American public, in the age of AI. The five principles offer guidance on the design, use, and deployment of AI, encompassing: automated systems; safe and effective systems; algorithmic discrimination protections; data privacy; notice, explanation, and human alternatives, together with consideration, and fallback [19].

Several important aspects of trust receive attention: management, in the form of stewardship; service, in terms of independence, genuine unfiltered access to the whole system; trustworthiness in the design, development and use and evaluation of AI products and services; innovation, in approaches to ensure trustworthiness, accuracy and explainability [20]; cyberspace [21], should be secure and trustworthy; bias (systemic, statistical and human), should not chip away at public trust; data brokers should be prevented from breeding corrosive distrust; public understanding and knowledge should be fostered through better explainability, to allow humans to appropriately trust and effectively manage the emerging generation of AI partners [20], addressing opaque decision making processes which result in a lack of public trust; and finally, recognizing the importance of placing trust in people and not technology [19].

Although these key approaches provide a basis for an assessment of trust, what appears not to be considered is a posture of zero trust, discussed next.

This paper suggests that a posture based on zero trust could be harnessed [13] to operationalize the US Government calls for the appropriate trust and effective management of emerging AI partners [19]. As highlighted, current approaches are based on presumptive trust [13] which can result in challenges such as unknown cyber security threats and or breaches. Adopting zero trust as a risk-based approach could arguably overcome these current limitations and challenges.

Implementing zero trust relies on an initial assessment of identity whether this is a an individual, or a device, service, or software [13], [15]. Once the identity has been verified trust can be granted, on a least privilege principle [14]. A posture of zero trust also involves continuous monitoring and verification [13]. Deploying zero trust could achieve the goal of securing a trustworthy cyberspace [19]. The operations of IVAs for cyber security in the context of zero trust are considered next.

This paper suggests that responsible stewardship, trustworthiness, and improved understanding [19] of IVAs could be demonstrated through implementing zero trust [13]. Returning to Table 1, above, all activities in the trust throughput model [10],[11], are reliant on the trustor's propensity. In zero trust, propensity is based on no presumptive trust [13],[14],[15],[20]. This approach compares with trust, where propensity is viewed as multidimensional and presumptive, implicitly trustful [10]. Given the viewpoint of zero trust, the antecedents of trust, involving both a trustor's ability to trust, and their experiences of trust [11] are dialed into a zero trust posture [20]. Such a position serves as a positive reinforcement to the posture of zero trust [20].

In the next stage of the throughput model, the assessment of trust, in this example, for zero trust, the propensity is again set as no presumptive trust [20]. In the assessment of ABI, as all traffic on the network is viewed as hostile and the state of the network is founded on a view of trust as a vulnerability [13], [14], [15] zero trust is once more reinforced [20]. In zero trust, if the assessment of identity (of device or service in the case of IVA) is verified, then confidence may be gained in the user and trust may be gained [14], [15]. In the next phase, in the trust model, the trustor's actions include assessing risk, accepting vulnerability, and taking risks [10]. In zero trust policy and enforcement decisions based are on authentication and authorization- these occur with no acceptance of risk or vulnerability [14],[15],[20].

Summing up, the findings from the approach presented here could provide steps that could overcome cyber security challenges and help inform judgments in policy decision and policy enforcement policy (PDP/ PEP) judgements [14],[16].

6 Concluding Remarks

This paper has examined IVAs, the problems of trust, zero trust and AI and reflections on current policy at the intersect between AI, trust, and cyber security have been presented. The contribution of the paper is to argue for the adoption of zero trust in the operations of AI and IVAs deployed for cyber security along with the elaboration of a new approach to assessing and evaluating zero trust. Suggested benefits through adopting these approaches together with promising avenues for future research and implications are discussed next.

This paper has demonstrated that leveraging well-established trust theory from organization studies and conflict resolution studies allows progress to be made in improving the steps in decision-making for policy and policy enforcement [14],[18]. Through this approach, users can benefit from transparency and understanding helping to fulfill the requirements for Trustworthy AI [20] and achieving the goals of the US 2022, AI Bill of Rights [21].

Promising avenues for future research include further study of decision making processes in policy development and in enforcement as zero trust is actioned. Further development of the conceptual model as presented could help practitioners better understand the key issues involved in building confidence in, and effectively assessing, managing, and monitoring our future generations of AI partners [20].

In conclusion, important implications for practitioners, policymakers and academics presented in this paper include an argument for the implementation of zero trust protocols and a process for implementation. It is hoped that scholars, practitioners, and policy makers will take up this call for further study and development.

Acknowledgements

Many thanks to colleagues for helpful discussions both online and in person. Thank you to the anonymous reviewers for their comments and observations that have improved this work.

References

1. ISC2 2022, Cyber security workforce study 2022, https://www.isc2.org//-/media/ISC2/Research/2022-WorkForce-Study/ISC2-Cybersecurity-Workforce-Study.ashx, last accessed, 2022/October/29.
2. Kaul, V., Enslin, S. and Gross, S., A.: History of artificial intelligence in medicine, Gastrointestinal Endoscopy, 92(4), 807-812, (2020), ISSN 0016-5107.
3. Adamopoulou, E. and Moussiades, L.: An Overview of Chatbot Technology. Artificial Intelligence Applications and Innovations, 584, 373 – 383 (2020).
4. Schmidt, B., Borrison, R., Cohen, A., Dix, M., Gärtler, M., Hollender, M. and Siddharthan, S.: Industrial virtual assistants: Challenges and opportunities. In Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, 794-801, (2018).
5. Chung, H., Iorga, M., Voas, J. and Lee, S.: Alexa, Can I Trust You? Computer 50(9), 100-104 (2017).
6. The Guardian, https://www.theguardian.com/technology/2022/oct/11/typos-and-shutdowns-robot-gives-evidence-to-lords-committee, last accessed, 2022/October/31.
7. Casilli, A., A.: Against the hypothesis of “the end of private life.” Revue française des sciences de l'information et de la communication. [in English], 3, (2013), https://journals.openedition.org/rfsic/630, last accessed, 2022/October/30.

8. Forbes, https://www.forbes.com/sites/jeanbaptiste/2019/07/30/confirmed-apple-caught-in-siri-privacy-scandal-let-contractors-listen-to-private-voice-recordings/?/, last accessed, 2022/October/30.
9. Fulmer, A. and Gelfand, M.: At what level (and in whom) we trust: Trust across multiple organizational levels. Journal of Management, 38(4),1167-1230 (2012).
10. Mayer, R., Davis J. and F. Schoorman, F.: An integrative model of organizational trust. Academy of Management Review, 20(3),709-734 (1995).
11. Deutsch, M.: Trust and suspicion. The Journal of Conflict Resolution, 2(4), pp. 265–79 (1958).
12. Mcknight, D. H., Carter, M., Thatcher, J., B. and Clay, P.: Trust in a specific technology: an investigation of its components and measures. ACM Transactions on management information systems, 2(2), 1-25 (2011).
13. Kindervag, J.: No more chewy centers: Introducing the zero trust model of information security. Forrester Research, Sept, 14, updated Sept, 17, (2010) https://media.paloaltonetworks.com/documents/Forrester-No-More-Chewy-Centers.pdf, last accessed, 2022/December/12.
14. S. H., S.: Zero trust architecture design principles. National Cyber Security Centre (NCSC), 20/Nov/20219. [online]. https://www.ncsc.gov.uk/blog-post/zero-trust-architecture-designprinciples, last accessed, 2022/October/30.
15. The White House, https://www.whitehouse.gov/?s=zero+trust, last accessed, 2022/Oct/31.
16. Rose, S., Borchert, O., Mitchell, A. and Connelly, S.: Zero trust architecture, NIST special publication 888-207, NIST, Aug/2020. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800207.pdf, last accessed, 2022/October/30.
17. Wylde, A.: Zero trust: never trust always verify. In: 7th International conference on Cyber Security for Trustworthy and Transparent Artificial Intelligence (CYBER SA 2021, IEEE), 1-4 (2021).
18. The White House, https://www.whitehouse.gov/ostp/news-updates/2022/10/04/blueprint-for-an-ai-bill-of-rightsa-vision-for-protecting-our-civil-rights-in-the-algorithmic-age, last accessed, 2022/October/31.
19. The White House, https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf/, last accessed, 2022/October/31.
20. NIST, https://www.nist.gov/system/files/documents/2022/03/17/AI-RMF-1stdraft.pdf
21. NSF, https://www.nsf.gov/pubs/2022/nsf22517/nsf22517.pdf