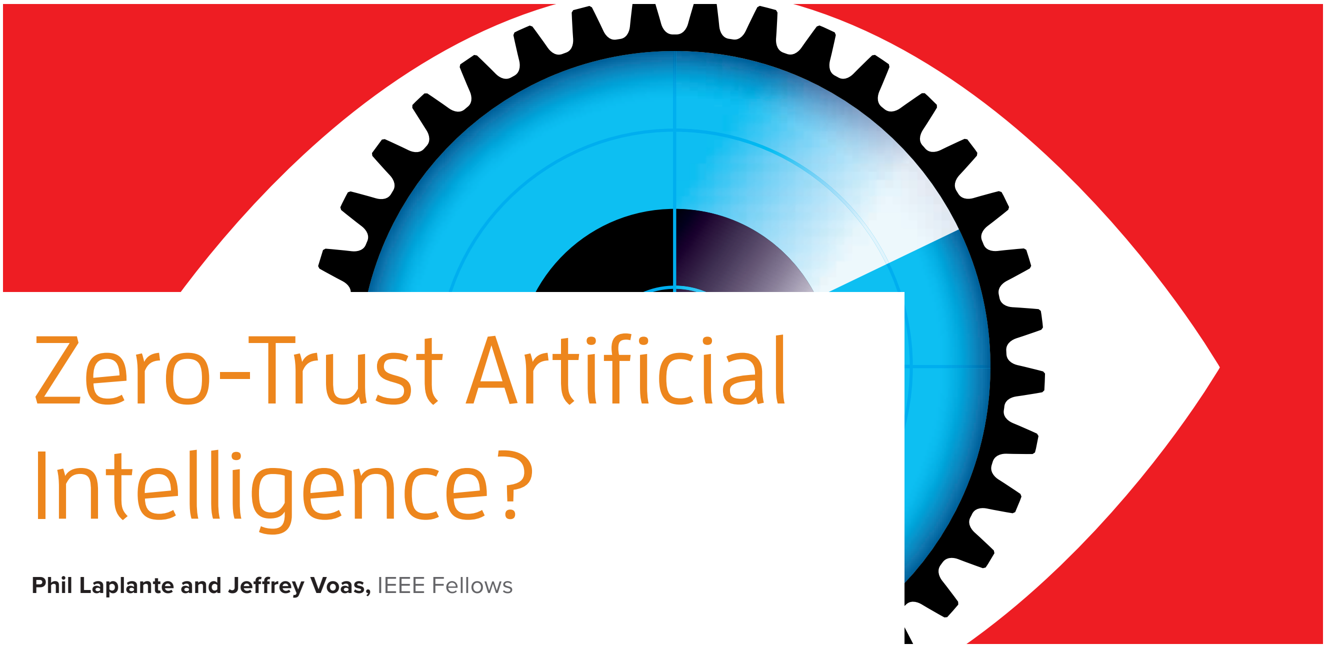


Zero-Trust Artificial Intelligence?

Phil Laplante and Jeffrey Voas, IEEE Fellows

Any critical artificial intelligence (AI)-based product or service should be continuously questioned and evaluated. This suggests a "zero-trust" or "trust, but verify" approach to AI.

Zero trust is a user philosophy. Absolute (or complete) trust sits at the other end of the trust spectrum; it's a producer's goal. The degree of achieved trust lies between them.[5] The zero-trust security architecture was introduced by Forrester Research in 2010.[3] "Zero trust is a cybersecurity paradigm focused on resource protection and the premise that trust is never granted implicitly but must be continually evaluated."[4] Here, we note that the artificial intelligence (AI) and cybersecurity communities have one noticeable commonality: few people understand AI and cybersecurity well. Hence, does "zero-trust AI" make sense? If not, then why do we need explainable AI?

Trustworthiness can be coarsely viewed as the degree to which a person has confidence that a service or product will behave as advertised, promised, and intended. Trust in an AI-based system can be described by the following three factors:

1. *Ability*: the capability of the AI system to do a specific task robustly, safely, reliably, and so on
2. *Integrity*: the assurance that information will not be manipulated in a malicious way by the AI system
3. *Benevolence*: the extent to which the AI system is believed to do good or where the "do no harm" principle is respected.[1,2]

We already have experience with AI in pedestrian systems, for example, recommender systems, home automation, entertainment systems, toys, and games. When AI-based products and services have the potential to cause physical and financial harm, that is, critical systems, trust becomes more relevant. We probably have less experience, but are familiar with, AI-based autonomous driving systems, and we are aware that AI controls and supports various critical infrastructures. Are you less trusting of these? For example, riding in a completely autonomous vehicle, undergoing unsupervised robotic surgery, and having AI adjudicate your murder trial means trusting your life to the technology's efficacy.

Consumer product and service companies often loosely include the word *trust* (or imply it) in their slogans and descriptions of their offerings. Imagine such slogans applied to AI-based offerings. In *The Gift of Fear: Survival Signals That Protect Us From Violence*, security expert Gavin de

Digital Object Identifier 10.1109/MC.2021.3126526
Date of current version: 14 February 2022

EDITOR IN CHIEF JEFFREY VOAS
IEEE Fellow; j.voas@ieee.org

Becker tells us that in dangerous situations, we should trust our instincts even when it might defy convention and logic. Our brains are exquisite pattern-matching machines that recognize danger as a matter of survival. Aside from playing chess and special, contrived situations, we probably should trust our brain's pattern-matching ability over AI, especially when the consequences could present danger. It is in those circumstances when our

IN THIS ISSUE

Computer regularly receives submissions on the topics of forensics, diagnostics, and tracing. Although these topics are not tightly coupled, I believe the five articles presented here come together nicely as a cohesive issue because in essence, these topics reveal "something" about what is going on in the internals of a system or ecosystem. This February 2022 issue features five articles that have been waiting to be published, and I'm pleased to finally release them.

In "Multilayered Diagnostics for Smart Cities," the authors discuss how smart cities employ technology to improve traffic patterns, energy distribution, air quality, and so on. They discuss how health care, education, culture, and shopping can all be integrated into a smart city while warning us about the need to consider cybersecurity and create the mitigating countermeasures against cyberattacks on such cities. The article examines smart city security threats from a multilayer perspective and offers a summary of attack scenarios and threat countermeasures.

In "Discovering Opioid Use Patterns From Social Media for Relapse Prevention," the authors study social media for communication and behavioral patterns of people with opioid use disorder (OUD). The article demonstrates how information derived from common activities such as online social networking might lead to better prediction and evaluation, ultimately preventing drug relapses. Through their multidisciplinary and novel analytic perspective, the authors characterize opioid addiction behavior patterns by analyzing opioid groups from https://www.reddit.com/, including modeling online discussion topics, analyzing text co-occurrence and correlations, and identifying emotional states of people with OUD. The article offers innovative ways to use information from online social media to create technologies to assist in relapse prevention.

In "Everything You Always Wanted to Know About Embedded Trace," the authors argue that intrusive software instrumentation and breakpoint-based debugging may not be the best options for observing operational system internals as they create complicated test flows and debugging procedures. This article proposes that embedded trace technology may be a technical answer to the observability conundrum that occurs in modern embedded computing systems. This article proposes embedded trace as an essential technology for testing and debugging toolboxes and highlights its capabilities, limitations, and opportunities.

In "A Multilevel Collective Framework for Internet of Things Digital Forensic Investigation," the authors focus on investigating crimes committed with or against Internet of Things (IoT) devices. The article's premise is that as sensor technology continues to become widely available, a need is created for intelligent and adaptable forensic models to investigate IoT-related crimes. IoT forensics collects and processes footprints from sources such as radio-frequency ID tags, smart devices, and cloud storage. The authors propose a new forensic investigation framework, MCF2I, which consists of 1) identification of IoT evidence sources at different levels and 2) multiphased investigation processes that coordinate different roles and tasks. The authors evaluate their approach on 32 users, including crime scene investigators and law enforcement officers, to show the success of their framework.

In "Privacy Guarantees of Bluetooth Low Energy Contact Tracing: A Case Study on COVIDWISE," the authors discuss how Google and Apple jointly introduced a digital contact tracing technology along with an application programming interface called *exposure notification* to help health organizations and governments perform contact tracing. The article examines and analyzes the security, privacy, and reliability of this new technology using 1) actual and typical case studies and 2) realistic use cases. Their experimental analysis validates the properties of the system in hopes of reducing fears of adopting exposure notification technology.

I hope you enjoy this issue.

—Jeffrey Voas, Editor in Chief

Digital Object Identifier 10.1109/MC.2021.3126527
Date of current version: 14 February 2022

zero-trust brains "kick in," and we should listen to them.

Is a zero-trust AI mentality always necessary? No. But we are suggesting that you should extend this instinctual, self-preserving distrust to AI. And we suggest that any critical AI-based product or service should be continuously questioned and evaluated; however, we acknowledge that there will be an overhead cost in doing so. Finally, is our question of zero-trust AI even worthy of discussion when AI is already ubiquitous and embedded in almost everything we rely on? You can't buy a new car and strip out the sensors and processors and expect it to work. So, maybe AI is simply a new, hidden, and unavoidable risk to life, devoid of opt-out options. Something to think about. ◼

DISCLAIMER

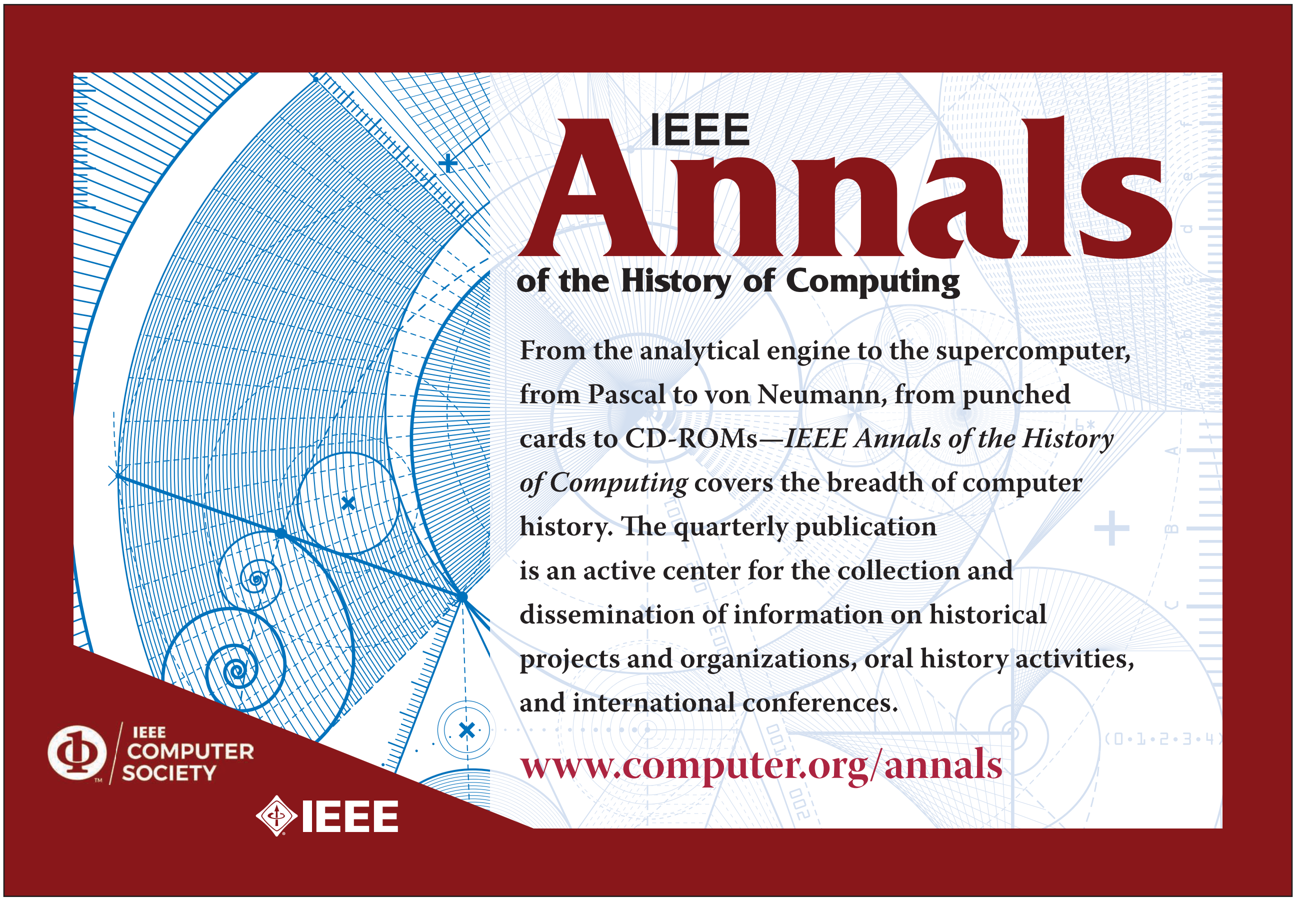
The authors are responsible for the content in this article. The opinions expressed are their own.

REFERENCES

1. *Information Technology — Artificial Intelligence — Overview of Trustworthiness in Artificial Intelligence*, ISO/IEC TR 24028:2020.
2. T. Zielke, "Is artificial intelligence ready for standardization," in *Proc. Eur. Conf. Softw. Process Improvement*, 2020, pp. 259–274, doi: 10.1007/978-3-030-56441-4_19.
3. "Next-generation access and zero trust," Forrester, 2018. https://go.forrester.com/blogs/next-generation-access-and-zero-trust/
4. S. Rose, O. Borchert, S. Mitchell, and S. Connelly, "Zero trust architecture," NIST, Gaithersburg, MD, USA, NIST Special Publication 800-207, 2020.
5. K. Miller, J. Voas, and P. Laplante, "In trust we trust," *Computer*, vol. 43, no. 10, pp. 85–87, 2010, doi: 10.1109/MC.2010.289.

PHIL LAPLANTE, Malvern, Pennsylvania, USA, is an associate editor in chief of *Computer*. He is a Fellow of IEEE. Contact him at plaplante@psu.edu.

JEFFREY VOAS, Gaithersburg, Maryland, USA, is the editor in chief of *Computer*. He is a Fellow of IEEE. Contact him at j.voas@ieee.org.



Digital Object Identifier 10.1109/MC.2022.3142830