AI for Cyberbiosecurity in Water Systems—A Survey

Daniel Sobien, Mehmet O. Yardimci, Minh B. T. Nguyen, Wan-Yi Mao, Vinita Fordham, Abdul Rahman, Susan Duncan, and Feras A. Batarseh

Abstract The use of Artificial Intelligence (AI) is growing in areas where decisions and consequences have high-stakes such as larger scale software, critical infrastructure, and real-time systems. This transition in recent years has been accompanied by the growth of research in AI assurance in fields such as ethical, explainable, and trustworthy AI. In this work, we survey the literature to find the state of AI assurance for cyberbiosecurity systems as they exist now, particularly for water and agricultural supply systems; future directions are also presented. We focus on papers at the intersection of cyberbiosecurity, AI assurance, and water/agricultural supply systems, discuss how assurance techniques improve these systems, and provide pointers for future research into the application of AI for the cyberbiosecurity field. Current cyberbiosecurity solutions do not focus much on AI, but existing AI solutions for water supply and cyber or Cyber-Physical Systems (CPS) exist and can

D. Sobien
Hume Center for National Security and Technology, Virginia Tech, Arlington, VA, USA
e-mail: sdan8@vt.edu

M. O. Yardimci · W.-Y. Mao
Department of Computer Science, Virginia Tech, Blacksburg, VA, USA
e-mail: oguzy@vt.edu; wanyi@vt.edu

M. B. T. Nguyen
Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA, USA
e-mail: mnguyen0226@vt.edu

V. Fordham · A. Rahman
Deloitte Touche Tohmatsu Limited, Arlington, VA, USA
e-mail: vfordham@deloitte.com; abdulrahman@deloitte.com

S. Duncan
College of Agriculture and Life Sciences, Virginia Tech, Blacksburg, VA, USA
e-mail: duncans@vt.edu

F. A. Batarseh (✉)
Department of Biological Systems Engineering, Virginia Tech, Arlington, VA, USA
e-mail: batarseh@vt.edu

D. Greenbaum (ed.), *Cyberbiosecurity*, https://doi.org/10.1007/978-3-031-26034-6_13

be applied to benefit cyberbiosecurity. The inclusion of AI assurances help alleviate issues of applying AI to high-stakes human-centered infrastructure.

CCS Concepts Computing methodologies → Artificial intelligence, Security and privacy, General and reference → Cross-computing tools and techniques, Computer systems organization → Embedded and Cyber-physical systems

Keywords Cyberbiosecurity · AI assurance · Water supply systems

1 Introduction

The deployment of AI is outpacing the adoption of assurances that commit to its responsible use as policies and regulations lag behind. Assurances validate AI systems to assess the risk of failure, misuse, and even abuse, helping establish the trust needed for the adoption of AI. The risks of AI in infrastructure (e.g., agricultural supply chains, biological systems, and water supply systems) are significant, potentially affecting millions of citizens and resulting in loss of life, well-being, and economic opportunity.

For example, take a city-wide water distribution system that pumps in water from a reservoir and ensures every citizen has equal access to drinkable water. Imagine the city adopts an AI system that predicts demand and supplies regions of the system as needed. The system works fine to start, but years later it is not properly validated after new pumps are installed, so the sensor data changes and no longer predicts accurately. As a result there are large swaths of the city that are no longer receiving drinking water because the system forecasts are off. Or, maybe the system was trained with bias data because poorer neighborhoods had less data collected, so the system favors keeping the water supply greater for affluent regions resulting in poorer regions having intermittent supply issues.

For this water supply AI system to work properly assurances must validate outcomes are correct, fair, and that users can understand why the system has made its decisions. These concepts form the basis of AI assurance, which details the broad ways of verifying and validating AI systems, much the same way that traditional programming software (i.e., not machine learning) is verified and validated during its development process [1]. AI assurance applied during development would help avoid the mentioned issues of robustness and bias.

Water supply systems are a form of CPS, as physical sensors, pumps, and tanks act as data collectors to track the flow of water and relay data to a central computer. This data processing exposes the water supply to cyber-attacks. Additionally, water supply systems are part of the bioeconomy (the supply chain infrastructure that is tied to critical commodities like food, water, and medicine) meaning any impact to the system can have an effect on the livelihood of thousands or millions of people. The imagined *water supply AI* not only ensures proper water distribution, but there are additional security concerns, moving it into the relatively new realm

of cyberbiosecurity, which is a discipline at the intersection of life science and information technology (IT) [2]. Cyberbiosecurity is defined in greater detail in Sect. 1.1.

Existing cyberbiosecurity research mostly focuses on the IT side of biology, or cybersecurity for biology labs and databases is a succinct way to put it. The cyberbiosecurity field, however, is lacking much research in applied AI for supply chain infrastructure, as most papers only identify vulnerabilities and propose high-level frameworks for addressing them. Our goal for this survey is to find papers at the intersections of cyberbiosecurity, AI assurance, and water and food supply systems and connect that to the bioeconomy. Our work searches for and discusses the applications of AI assurance to existing solutions within the cybersecurity and CPS to help ensure the proper function of cyberbiosecurity-related systems.

1.1 Relevant Terminology and Definitions

Proper use of AI assurances verifies and validates the outputs of those systems, convincing users that they are reliable. AI assurance codifies the process, so when changes occur to the water supply system, validation can be re-run to satisfy the AI is working properly or needs to be retrained. Definitions are intentionally broad in order to apply them to a wider range of applications. From Batarseh et al. [1], AI assurance is defined as:

> A process that is applied at all stages of the AI engineering lifecycle ensuring that any intelligent system is producing outcomes that are valid, verified, data-driven, trustworthy and explainable to a layman, ethical in the context of its deployment, unbiased in its learning, and fair to its users.

The importance of AI assurance is that it applies a process to all stages of the AI lifecycle, from the start of development all the way through deployment. Assurances are not merely tests of AI to check some boxes that it is okay to use. In order to trust the AI is working properly engineers need to validate it meets all the criteria of assurance:

- Ethical—the AI system can make "right" decisions that benefit the people impacted and not just the people in power of the technology [3].
- Fair—the AI system makes decisions without considering demographics, backgrounds, affiliations, or individual preferences (i.e., does not inherently value some citizens over others).
- Safe—the AI system ensures the life and well-being of those who are using it and impacted by it.
- Explainable—the AI system can explain, or be interpreted, to understand why it came to a decision or how the algorithm works.
- Secure—the AI system can prevent or mitigate attacks or other threats to the proper operation of the system.
- Trustworthy—users have confidence the AI system works properly.

For infrastructure systems in the bioeconomy, AI must be ethical to make the right decisions, safe to protect users it potentially impacts, explainable so humans can understand it, fair in the decisions it makes, trustworthy so we have confidence in its abilities, and secure to prevent cyber-attacks and threats.

The bioeconomy refers to the sector of the economy that relates to research or innovations in the life and biological sciences and fields related to biotechnology [2, 4–6]. This sector grows as progress continues in technology relating to computing and information sciences [7], including most crop production, especially as big data, AI, and machine learning become more involved for enhancing land use and water management via precision farming [4]. As the bioeconomy grows, cyber threats against it increase and require mitigation to safeguard investments in the bioeconomy [8].

Richardson et al. [2] described cyberbiosecurity as the intersection of IT and life sciences, but Duncan et al. [9] specified it further as the intersection of cybersecurity, cyber-physical security, and biosecurity. Each discipline with its own existing challenges and new vulnerabilities appearing where they overlap.

By its nature, cyberbiosecurity is grounded in IT and with that brings the risk of cyber-attacks. This is the traditional realm of cybersecurity, or the shielding of computer networks and information from damage, exploitation, and unauthorized use [10–15]. Linking any computer system to a network increases risk. This is compounded in the bioeconomy as more remote monitoring and controlling is added to existing physical infrastructure, because of this interaction of cyber and physical the security needs "safety and reliability requirements qualitatively different from those in general-purpose computing." [16]. A CPS integrates digital computing and physical processes, where a network monitors and controls a physical system via sensors and actuators, to interact with the real world [16, 17]. Communication and networking multiple devices is important because the components are often disparate and there is a back and forth of physical processes affecting the computer and vice versa, but this opens new vulnerabilities [16, 17].

The third aspect of cyberbiosecurity moves fully into the physical space for securing biological systems. Biosecurity is the protection of any form of life from the threat of disease and pests, including the protection of agriculture and food, or simply put the "re-branding of the centuries-old battle with disease" [18–20]. This includes threats that are natural, such as livestock and crop diseases, or intentional attacks, such as the deliberate use of smallpox and anthrax weapons [18]. The incorporation of biosecurity in the realm of cybersecurity and cyber-physical security is what sets cyberbiosecurity apart.

Traditional cyber-attacks are not necessary to impact biological systems, because there are physical, biological interactions outside the computer systems. We need to ensure that the biological aspects are operating properly, be it from natural causes (diseases, pests, etc.) or intentional cyber and physical attacks. There are three layers of interactions to protect: the cyber, the interactions of cyber and physical, and the biological.

Included in these biological systems are water supply systems, which can refer to distribution, treatment, agricultural, or storm water systems. Distribution systems

control the transport and delivery of water through a network of pipes and pumps to ensure consistent supply, they are focused on the logistics of water transportation and storage. Treatment systems take raw or wastewater, unsafe for humans or the environment, and through a series of chemical and biological processing, filtering, and sanitizing produce either safe drinking water or water that can be released into the environment. Agricultural water systems focus on the distribution of water to crops and livestock. Unlike distribution systems, this water does not have to be safe for human drinking, but it must ensure the production of food for human use. This also closely ties agricultural water systems to food supply systems. Finally, storm water systems deal with the drainage of runoff water to prevent flooding or contamination of other water systems from the pollutants that it picks up.

These systems allow for the automation of critical infrastructure by adding more technology for monitoring and controlling human and agricultural water use. These water and food systems are not only cyber-physical but also biological as well. Their proper functioning is required for human livelihood, either through the supply of safe water or the growth of adequate food supplies. Water and food systems are cyber-physical and bio-infrastructure systems that are open to attacks (cyber and physical) and anomalies (such as maintenance issues, severe weather, sensor or equipment breakdowns).

Going back to our hypothetical city-wide water distribution system. If it were attacked by a bad actor who wanted to poison the water, they could give commands to add too much of a chemical or too little of a cleaning agent that would result in undrinkable water. In fact, there was an attack in 2021 on a Tampa, Florida water supply system where attackers increased the levels of lye in the water by 110 times before they were stopped [21]. We discuss this example further in Sect. 5.4, but it serves as a great example of the cyberbiosecurity threats to water supply systems. Threats can combine unauthorized access of computer systems to control physical processes; in the Tampa case, the lye controllers pose a biological threat to everyone that relies on the system for safe drinking water. The next section introduces the inclusion and exclusion criteria of the papers surveyed.

1.2 Description of Included Articles

In this survey, we used multiple online repositories and research paper search engines to find relevant papers on the topics of cyberbiosecurity, AI assurance, and water supply systems. Our focus was to find peer-reviewed papers at the intersection of two or more topics. We include papers from journals, conference proceedings, dissertations, books and book chapters, and industry white papers published from **2000** through **April 2022**. A complete repository of papers included in this study can be found here: https://github.com/AI-VTRC/CyberbiosecuritySurveyPaper.

Key search terms included the following to find papers:

- Cyberbiosecurity; Cyber-Biosecurity; Biocybersecurity; Bio-Cybersecurity

- Water Supply System; Water Distribution System; Water Treatment System; Water System
- AI Assurance (see assurance list in Sect. 1.1)
- Artificial Intelligence

Because cyberbiosecurity is a new research field, we kept search criteria as broad as possible to include enough papers for a survey. Some focus on the medical fields, but we tried to find relevant discussions that could apply to AI assurance or water supply systems as much as possible. Some focus just on the concept of cyberbiosecurity in general, but we focus on how best to apply the concept to AI assurance and water supply systems.

2 Survey Landscape

The papers surveyed for this research included publications between 2000 and 2022 (as of April 2022), but most are from 2016 onward. Figure 1 shows a histogram by publication year, and until 2016 there was not more than three publications per year that covered cyberbiosecurity, water systems, and AI assurance. There is a steady trend upward for the count of publications, and as cyberbiosecurity and AI assurance

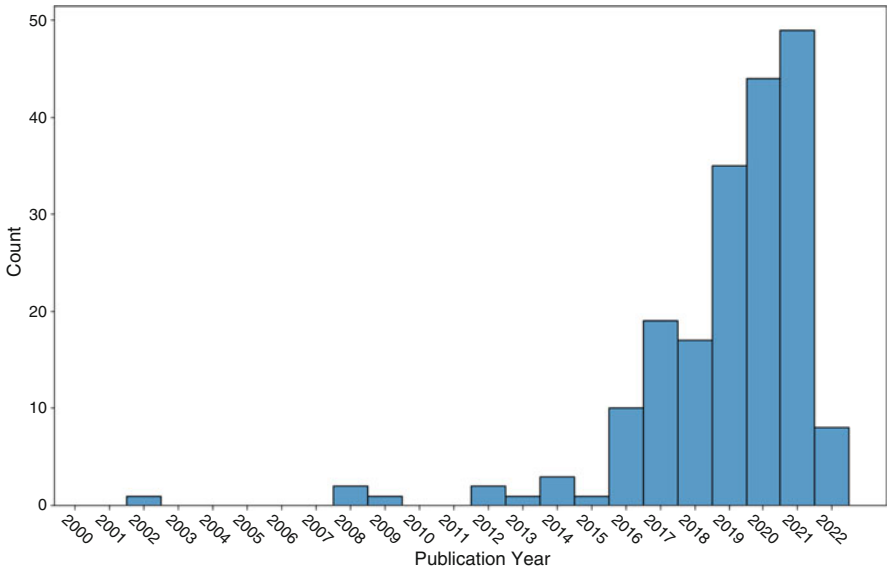

Fig. 1 Count of the number of publications by year that were used in this survey

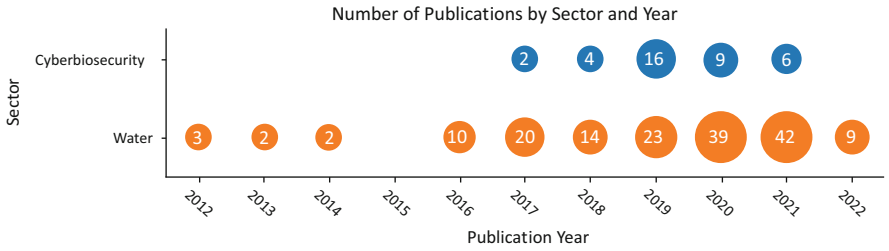

Fig. 2 The count of publications by year for the sectors of cyberbiosecurity and water supply (either water treatment or water distribution) systems. Papers are not confined to a single sector, and some are counted both as cyberbiosecurity and water supply papers. Most papers published since 2012, so older publications omitted from this figure

research continues to grow we expect the number of publications to continue to grow each year.

Figure 2 shows the breakdown of publications by cyberbiosecurity and water sectors. Publications on water systems had a low but steady trend from the early 2000s until about 2017 when they increased and held since. The year 2017 was also when the cyberbiosecurity term started showing in the scientific literature, and there is a sharp peak in 2019 before cyberbiosecurity publications return to a more steady pace.

We break down the AI assurance publications by assurance pillars in Fig. 3. Here, a majority of the papers deal with safe and trustworthy AI, especially just before the term of cyberbiosecurity starts showing in 2017. As AI becomes more popular, especially with deep learning (since 2015), we see an increase in publications for all the pillars of AI assurance.

Figure 4 shows a citation graph we created using Citation Gecko.[1] The yellow nodes are surveyed papers, gray nodes are other papers which cite our surveyed papers, and edges (lines that connect the nodes) are the citation link between two papers. The cyberbiosecurity literature is relatively disjointed from the literature on water supply systems and attack/anomaly detection. Most of the AI assurance papers remain independent in this view from each other and other sectors, with the exception of some trustworthy AI papers that form a small network. This graph shows the relative separation of the cyberbiosecurity literature from water system security and attack/anomaly detection (which includes secure AI). There is one citation chain from cyberbiosecurity to water system security via Mueller [22], Schmale III et al. [23], Moyer et al. [24], and Housh and Ohar [25]. (note that Moyer et al. [24] is the oldest link in that chain.)

[1] https://www.citationgecko.com/.

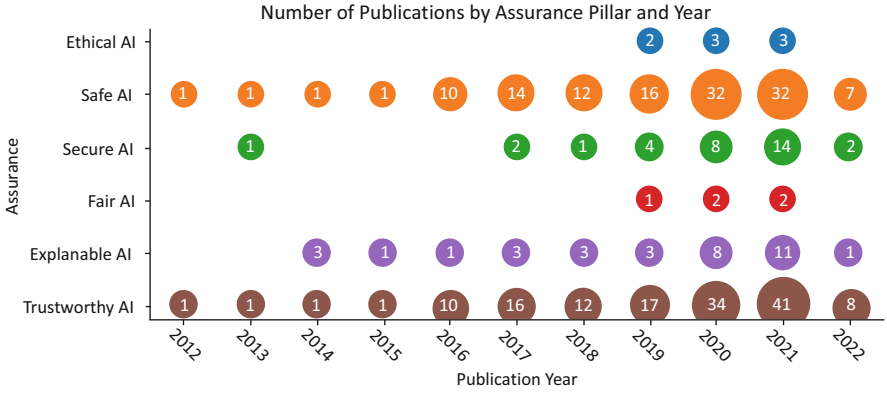

Fig. 3 The count of publications by year for the pillars of AI assurance. Papers are not confined to a single pillar, and some are counted for multiple. Most papers published since 2012, so older publications omitted from this figure

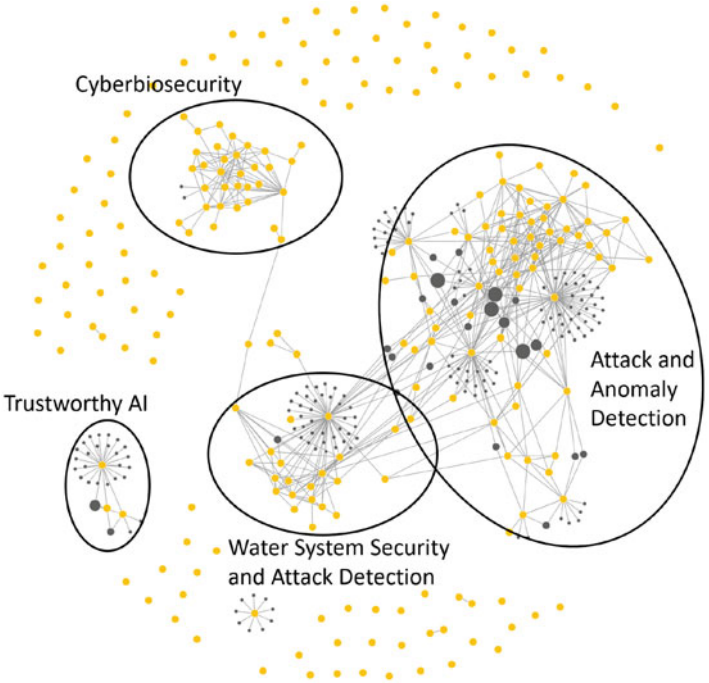

Fig. 4 Connected citation graph of the papers survey for this work. Yellow nodes are surveyed papers, gray nodes are other cited papers, and edges represent a citation between two papers. The cyberbiosecurity literature is relatively disjointed from the literature on water supply systems, AI assurance, and attack/anomaly detection. Graph generated using and courtesy of CitationGecko https://www.citationgecko.com/

3 AI Assurances for Cyberbiosecurity

In the introduction section, we described cyberbiosecurity as the intersection of life sciences and IT, and to be a little more specific it is the intersection of cybersecurity, cyber-physical security, and biosecurity [2, 9]. One of the best definitions we found is from Murch and DiEuliis [26], who defined cyberbiosecurity as the

> **understanding [of] the vulnerabilities** to unwanted surveillance, intrusions, and malicious and harmful activities which can occur within or **at the interfaces of commingled life and medical sciences, cyber, cyber-physical, supply chain and infrastructure systems**, and developing and instituting measures to prevent, protect against, mitigate, investigate and attribute such threats as it pertains to security, competitiveness, and resilience. (emphasis ours).

It is the vulnerabilities at the intersections of these cyber, physical, and biological systems that make cyberbiosecurity what it is, complex interactions between machines and biology that are open to disruption. This interaction creates unique vulnerabilities open to biological systems that make detection, attribution, and mitigation difficult in a timely manner [27]. Bernal et al. [28] recreated a Distributed Denial-of-Service (DDoS) attack using bacteria "engineered to act as biosensors" in a novel cyberbioattack, demonstrating the unique risks of the field and that traditional cybersecurity measures are not always adequate for cyberbiosecurity applications. The literature addresses these issues with a widespread call for action and collaboration—"We call for analyses and publications to fully scope cyberbiosecurity and identify a comprehensive strategy to establish the discipline's goals and objectives" [2] and others, as called out by [29] and seen in [26].

The purpose of our survey is to find how cyberbiosecurity intersects with AI assurance; there are applications that go beyond applying security to biological applications, and here we are interested in answering the question: what makes cyberbiosecurity different than cybersecurity for biology? It is the assurances a cyberbiosecurity system brings to the continuing function of the bioeconomy and relevant infrastructure. This is summed up well in the paper from Schmale III et al. [23], and while cyberbiosecurity is only mentioned briefly, the goal of the water supply system discussed is to ensure the safety of the drinking water from naturally occurring harmful algal blooms and cyber-attacks. Cyberbiosecurity "models must capture the physical dynamics of the system as well as the cyber-interconnections" [23].

Cyberbiosecurity systems that deal with supply chain and infrastructure systems have, or the potential to have, large impacts on the livelihood of people who rely on the system. All the residents of a city rely on its water distribution system to bring them water for drinking, cooking, and cleaning. A break down is not merely inconvenient but could be life-threatening, especially if the system is down for a long time or the water is contaminated. Even if AI is not considered for a cyberbiosecurity system, assurances are important to what cyberbiosecurity attempts to accomplish. AI brings an opportunity to add security or corrective actions in the event of any issues, and AI assurances validate their use for cyberbiosecurity applications. The

end goal of any assurance (AI or not) is validating and verifying a system is working properly, so people have trust and adopt that system for use.

Turning back to the example of a water distribution in a city, suppose an AI monitors the system for cyber-attacks or natural anomalies (e.g., low levels from draught, bacterial growth, broken equipment, etc.) and takes corrective actions. If the hypothetical water distribution AI meets all the criteria listed in Sect. 1.1, then there is assurance that it behaves in a way that benefits everyone it impacts (people in the city who rely on the system providing drinkable water on demand) and minimizes unintended consequences. There is also some assurance the AI mitigates issues or threats to the system that would endanger city residents.

All these AI assurances are relevant to cyberbiosecurity, especially the secure assurance because the objective of cyberbiosecurity is "understanding the vulnerabilities" and developing "measures to prevent, protect against, mitigate, investigate and attribute such threats as it pertains to security…" [26]. There is also the human side of cyberbiosecurity, Perakslis [30] included the field in their list of public interest technologies, which are technologies that focus on public good. Further emphasizing the need for assurances to validate any AI systems involved with cyberbiosecurity and help promote their adoption in cyberbiosecurity. AI systems need to be trustworthy and explainable so people want to use them knowing they can rely on them to operate correctly, and because cyberbiosecurity systems focus on biological systems, safety is a big issue in order to ensure people impacted are not threatened by AI making a wrong decision. Ethics and fairness are a large part of the safety assurance too, as AI needs to ensure it does not favor some people over others, that it is not designed to favor its developers and investors over everyone else. Ethics and fairness are ensuring equal safety for everyone impacted.

4 AI Assurances for Open-Source Water Supply Testbeds

Open-source information engages more researchers allowing them to build better tools, frameworks, and operational systems such as Git, PyTorch, or Linux. Similarly, open-source testbeds allow the community to contribute, propose, test, and improve upon ideas. Lack of real-world water and CPS datasets prevented significant research in security of these systems [31]. Data from real facilities cannot be shared for both security concerns and lack of accurate ground truth, so the availability of reliable, open-source water testbeds is critical for research. Open-source datasets also allow hands-on experience and training scenarios needed for collaboration and understanding the security requirements of these systems [32].

Assurances for water systems closely match those of cyberbiosecurity systems discussed in Sect. 3. The two major assurances are the safety of the water quality and the security of the system's operations. Explainability is another key assurance for water systems, so we can understand how the water and AI systems operate in order to ensure consistent and safe water supplies. This emphasizes the importance of open-source datasets to help the AI research community better understand the

operation of water systems and develop explainable and interpretable AI that is open to the water industry. Here we present some open-source water distribution and treatment system (as defined in Sect. 1.1) testbeds available to researchers across the world [33].

4.1 Secure Water Treatment (SWaT) Dataset

SWaT is a scaled down water treatment plant with real cyber and physical equipment to investigate cybersecurity research, which started in 2015 by Singapore University of Technology and Design [31]. The testbed consists of a six-stage water treatment process with modern-day components. The data collected from the testbed consists of eleven days of continuous operation, including seven days' worth of data under normal operation and four days' worth of data under attack. All network traffic, sensor, and actuator data was stored in the database.

4.2 Water Distribution (WADI) Dataset

Due to the success of the SWaT testbed, Singapore University of Technology and Design launched WADI in 2016 as an extension of SWaT to form a complete water treatment, storage, and distribution system [34]. Similar to SWaT, data collected for the WADI testbed consists of sixteen days of continuous operation, including fourteen days' worth of data under normal operation and three days with attack scenarios. All network traffic, sensor, and actuator data were collected.

4.3 Battle of the Attack Detection Algorithms (BATADAL) Dataset

The BATADAL dataset is not based on real-world data, though it is considered realistic since it was constructed using the de facto standard simulation tool for water distribution system modeling, namely the open-source Matlab software package EPANET [35]. EPANET is a Windows based software application for simulating and representing water distribution systems used world-wide by engineers and researches to design new water infrastructure, update existing water systems, and develop more efficient solutions to solve water quality problems. The BATADAL dataset was constructed for a competition to compare the performance of algorithms for the detection of cyber-attacks on water distribution systems. BATADAL simulates a fictional C-Town water distribution network, first introduced for the Battle of the Water Calibration Networks by Ostfeld et al. [36]. C-Town is based on a

real-world, medium-size network which contains 388 nodes, 429 pipes, 7 tanks, 11 pumps, and one actionable valve.

4.4 Modbus Penetration Testing Framework (Smod) Dataset

Laso et al. [37] created the Smod dataset was produced in 2017 to investigate how data and information quality estimation can detect anomalies and malicious acts in a CPS. The data were acquired using a cyber-physical subsystem consisting of liquid fuel or water containers, along with its automated control and data acquisition infrastructure. The data consist of temporal series representing five operational scenarios—normal, anomalies, breakdown, sabotages, and cyber-attacks—corresponding to fifteen different situations. To acquire the data, Laso et al. [37] used two tanks of different volumes for storage, one ultrasound depth sensor, four discrete sensors, and two pumps.

4.5 Digital Hydraulic Simulation (DHALSIM) Framework

DHALSIM is an upgraded framework of the BATADAL Framework, which uses the Water Network Tool for Resilience (WNTR) EPANET wrapper to simulate the behavior of the water distribution systems [38]. DHALSIM uses Mininet and MiniCPS to emulate the behavior of the Industrial Control System (ICS) controlling a water distribution system. This means that in addition to physical data, DHALSIM also provides network captures of the Programmable Logic Controller (PLCs), Supervisory Control And Data Acquisition (SCADA) server, and other network and industrial devices present in the system. Similar to BATADAL, DHALSIM can be integrated into a C-Town Network, using a Mininet network that connects the C-Town PLCs and SCADA servers through Local and Wide Area Networks (LANs and WANs). In DHALSIM, each ICS equipment is a Mininet node running a script that represents the behavior of such equipment. In the C-Town network PLCs have private Internet Protocol (IP) addresses and NAT and port forwarding is used to connect the LANs.

4.6 Datasets Comparison

Figure 5 compares the number of total citations (labeled “General Citations”) to the number of cyberbiosecurity citations (labeled “Cyberbiosecurity Citations”) for the five datasets above. We obtained the number of cyberbiosecurity citations and general citations by counting the numbers of papers citing these datasets in our survey and by the count of citations from Google Scholar, respectively. We see the

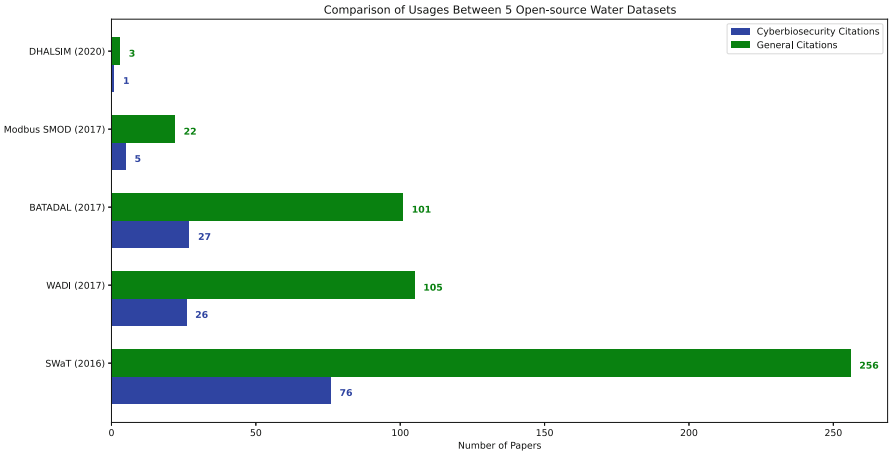

Fig. 5 Comparison between five open-sources water datasets in term of data usage

SWaT dataset is used the most, while DHALSIM dataset is used the least in both types of citations. This difference could be explained due to the early deployment of the SWaT dataset and the continuing collection and publishing of more data to that dataset by the University of Singapore in the years since its initial release. Although SMOD, BATADAL, and WADI are all water distribution systems published in 2017, the SMOD dataset is used significantly less. This could be explained by the scale of the datasets, specifically, both BATADAL and WADI simulate water distribution systems of large towns with multiple sensors, nodes, pipes, and a large recording time. On the other hand, SMOD only simulates a two-tank system, although SMOD is focused on different attack and anomaly scenarios than BATADAL and WADI. This shows that the research community prefers a dataset that can simulate a large scale, high quality real-world water distribution systems (WADI and BATADAL) and water treatment plant (SWaT) as benchmarks for model development.

5 AI Assurance Pillars

AI offers both opportunity and risk to cyberbiosecurity systems. It has the potential to detect and mitigate cybersecurity threats [2, 39–42], but at the same time offers an avenue for attacks [43–45], such as "poison" and "evasion" attacks on data or "inversion" attacks on AI models [43]. The current state of the cyberbiosecurity literature, however, focuses more on creating awareness and calls for collaboration to mitigate security threats rather than discussing the direct use of AI or AI assurance.

This supposition is not uniform, as Reed and Dunaway [40] praised the use of AI to "assist decision making… through the identification of cyberbiosecurity vulnera-

bilities and by providing recommendations for their elimination and/or mitigation." AI already brings a lot of benefit to the field of cyber and cyber-physical security, so the extension to cyberbiosecurity seems inevitable. However, with different physical, biological, and safety considerations required for cyberbiosecurity, there are no guarantees of success. This is where AI assurances come in to play a role, as they can help validate AI systems function as intended and aid in the responsible adoption of AI for the field of biology [1, 2, 46].

The multifaceted issues and solutions cyberbiosecurity systems face require interdisciplinary teams [47]. Solutions, therefore, cannot only be technical but require just as much of a human element [2, 47–49], and this is a more common topic in the surveyed papers than direct mentions of AI for cyberbiosecurity.

Assurances aid the adoption of AI by evaluating them for the benefit of humans and not because they make a solution more efficient, cheaper, or faster. The pillars of assurance are ethical, fair, safe, secure, explainable, and trustworthy. With the exception of secure, they are completely human focused. Clark et al. [48] claimed that cyber-defense is comprised of three aspects: technology, people, and physical protection and that these applications rely on people merging their knowledge rather than solely relying on automation. AI assurance is the way of merging the technological solutions of AI with the human values of the people within the cyberbiosecurity ecosystem. Aguilar et al. [49] argued a more holistic approach is required to solve the issues with the bioeconomy, one that includes "science, technology, economy, environmental issues, rural and industrial development, regulatory processes and social sciences."

5.1 Ethical and Fair AI

The most important question we can ask about AI is whether it works as intended or not. If not, how bad can the results be? And what kind of measures can we take in case of such a failure? In March of 2018, "an autonomous car operated by Uber—and with an emergency backup driver behind the wheel—struck and killed a woman on a street in Tempe, Arizona. It was believed to be the first pedestrian death associated with self-driving technology" [50]. This incident is a crucial example of when AI fails to make a safe decision. Although writing detailed contracts can legally reduce a manufacturer's liability, it might be morally unethical for the company to avoid legitimate liability.

With the growth of AI there are ethical and legal concerns regarding technology in areas, including how we can eliminate AI biases, ensure privacy, facilitate safety, and much more. AI should be made trustworthy, should be created and used with "an ethical purpose," and created to do good in society, but there are lots of questions that come up with AI and robots, such as if we "[assume that] the robots cannot be morally responsible—who will be responsible?" [51]. Furthermore, AI is already used in automated decision-making, and in high-stakes scenarios their decisions can be impactful. One issue with algorithmic decisions is bias, which can be "cognitive

biases of programmers," "unrepresentative datasets used for training," or "bias in the data used to make the decision" [51]. It is just as important to start with ethical considerations before AI is designed, let alone deployed, to ensure it is making fair and ethical decisions [51].

The concerns of inclusive, equitable, and correct decisions from AI are not solely left to industry, in fact it is gaining more ground in research from large tech companies and academics. The ambiguity of "fairer" decision-making systems, however, leaves fair AI as a broad open ended question without a real solution. Besides defining what "fair" means, researchers must deal with how to train systems for fair decisions or the fact that systems made fairer for one group can result in bias against another.

One of the most common reasons for biased results is the under-representation of certain groups within a dataset. Increasing the representation of that group, for example, oversampling a certain demographic in certain areas predominantly held another, may be a solution to rectifying the data. When it is not possible to modify or edit data, the objectivity of the decision-making process can be resolved by adjusting the AI algorithm. For algorithms that learn from discriminatory practices it is possible to change the internal weights in a way that makes decisions more neutral. It is also possible to modify the decisions of AI algorithms directly to create more equitable outcomes.

In some instances, it is not the lack of representation, but rather, the over-representation on certain groups that can created biased results. In such fairness related cases, openness in the development and deployment of AI is required [52–55].

In short, it is possible for AI technologies to be more equitable, but this requires the cooperation of different stakeholders and a lot of work. Arnold et al. [56] pointed out the importance of ethical decision-making while raising critical questions for every AI developer. The authors also refer to relevant answers for these questions from the literature, making this article serve as a guidebook for comprehensive AI assurance deployment.

Laplante et al. [57] investigated the causes that lead to unethical AI and its potential results. The authors saw the main reason as unbalanced or underrepresented data. [57] also emphasized the importance of ethical considerations for AI over its importance for classical software.

Zicari et al. [58] provided a framework to assess the trustworthiness of AI systems. The parameters the authors investigated include, but are not limited to, ethical and fair AI. The article provided a lifecycle to ensure ethics in AI decision-making. The authors emphasized the required absence of conflict for a reasonable assessment of ethical AI.

Grady et al. [59] proposed an epistemic, ethical analysis framework; as the name suggests, the authors proposed ways to detect and analyze ethical issues in cyber-physical infrastructures including, but not limited to, water treatment and distribution systems. The article investigated the importance of ethical decision-making and the roots of the problems in this topic.

Freeman et al. [46] proposed a framework to investigate AI using AI assurance metrics. The authors brought together many AI measures on common ground in this work, challenged the readers, and provided answers to these AI assurance problems.

Calvo et al. [60] investigated the algorithmic, environmental, and human impact assessment of AI systems. They proposed a measurement algorithm called Human Impact Assessment for Technology (HIAT) and discussed ways to build trust into the algorithm using this method.

5.2 Safe AI

One goal of cyberbiosecurity is ensuring the safety and well-being of those impacted by the system. This stems from the biosecurity aspect of the field [61] but naturally extends to any form of safety ensured by systems like water and food supply chains (and agriculture [62] as an aspect of these supply changes). The goal of the safe AI assurance is for AI to guarantee some level of safety to ensure the life and well-being of anyone impacted by the AI. These two forms merge to, as Mueller [22] described cyberbiosecurity, develop, validate, and implement safety measures.

Physical consequences, including harm to humans, are what separates cyber-biosecurity from most forms of technological security. Walsh and Streilein [43] pointed out that "a successful cyber intrusion within the bioeconomy may yield a result that causes physical harm, something generally associated with biosafety and biosecurity but not cybersecurity." Any interference with the bioeconomy has potential to harm, and while Walsh and Streilein [43] focused on illicit interference, this extends to unintentional interference as well. It is the ability for any cyberbiosecurity system to cause physical harm, intentional or otherwise, that safe AI and safety assurances need fortifying.

Water and food supply systems are a prime example of a cyberbiosecurity systems where safety is a priority. Quality and supply from the system impact everyone in a service region, and both are affected by natural anomalies (algal blooms, weather, draughts, and floods) or cyber-attacks. Water supply systems require constant monitoring and threat mitigating to ensure safety of the water quality and supply [23, 63–79]. On the other hand, food supply relies less on technological innovations, whereas water systems have standardized the use of SCADA systems [48], food supply and agriculture have seen a more limited and hesitant adoption of technology, especially for small-scale farmers [9]. A more standardized approach to tech adoption helps by "securely sharing and interpreting data across sectors and identifying cyberbiosecurity risks," ultimately improving food supply chains by designing "agricultural and food systems to better meet consumers' need and protection of life science data" [80]. Data privacy is also a concern any time personal health information may be involved with genomic databases with the potential for cyber-attacks on lab automation [81, 82].

We found in the literature that water and wastewater sectors vary greatly in size, complexity, organization, security protocols, available resources, and even in

imposed regulations [47, 48]. While the end goal of each water system is to supply clean water on demand, the approach each system takes is unique and requires different considerations, including adopting security measures specific to their organization [48]. This means that each system needs to take unique considerations to ensure to the quality of the water and consistency of the supply, posing a challenge to the field as a whole because standardized approaches to safety cannot be developed or relied on for all situations.

The bioeconomy, too, consists of large and complex systems that intertwine and connect, and it "harbors unique features that have to be more critically assessed for their potential to unintentionally cause harm to human health or environment" [22]. Water systems supply water to farms that impact agricultural production which in turn impacts food supplies to retails (grocery stores), prices, and the ag-economy. Any hiccup along the way can have unforeseen consequences. The complexity, however, makes it difficult for any one person, or even organization, to understand what consequences their actions have. This means that changes for the sake of mitigating external threats could lead to unintended consequences [39]. Cyberbiosecurity cannot focus solely on cybersecurity and attack detection or, as mentioned in the previous section, on monitoring natural phenomena as interference. We need to implement assurances to guarantee the safety of a system (e.g., quality of water or food for human consumption) at all times.

AI and other emerging technologies' reliance on data provides both benefit and potential harm. The concern of unintentional errors can arise in the data used for Safe AI. Caswell et al. [83] pointed out the potential issues of errors in biological databases, but the concern is applicable to any data-driven analysis in cyberbiosecurity. While referring to synthetic biology, Li et al. [84] emphasized that unintentional risks can lead to food scarcity despite the efforts of biosafety and biosecurity to provide more. Similar concerns for unintended consequences of dealing with biological data have been expressed in [84, 85]. As these technologies are implemented more into cyberbiosecurity systems (such as precision agriculture) more emphasis needs to be placed on quality assurance of the data and safety assurances for the final product.

5.3 Explainable AI

In the introduction section we defined explainable AI as AI that can "explain, or be interpreted, to understand why it came to a decision or how the algorithm works." Here, we expand this to include cyberbiosecurity systems in general because that is the environment the AI system operates in, the AI's behavior is dependent on the larger system, and the end user needs to understand both in order to operate the system correctly. Even if a cyberbiosecurity system does not incorporate AI, human understanding is crucial to its operation. Therefore, we expand the definition of explainability to include "the process of making complex systems human intelligible."

The literature surveyed often mentions the lack of training, understanding, and even awareness of cyberbiosecurity and cybersecurity risks as a vulnerability. This means a lack of knowledge and human understanding of threats, how to recognize them, and what to do about them is one of the biggest hurdles for the cyberbiosecurity field to overcome. Accordingly, a framework for making these complex systems understandable in order to avoid and mitigate risks is recommended. However, even in the biotechnology and cybersecurity realms "cyberbiosecurity is not well-known or understood" [86] and there is "a failure to recognize vulnerabilities" [40]. This lack of awareness is detrimental because cyberbiosecurity relies on understanding the vulnerabilities, threats, and risks to mitigate impacts [22, 26]. Even with the conventional cybersecurity approach, a "good cybersecurity plan is understanding the threat and establishing cybersecurity governance protocols" [47]. The mentioned approaches are not fully implemented or are done so inadequately resulting in "the failure of individuals to identify and address cybersecurity vulnerabilities" in cyberbiosecurity systems [40].

Part of this lack of awareness is from lack of education or training available in cyberbiosecurity [87]. Drape et al. [29] surveyed researchers from the agricultural sector attending a cyberbiosecurity workshop and found that no participants had cybersecurity training or resources, and attendees were uncertain about obtaining training or implementing solutions. Despite the research going into cyberbiosecurity vulnerabilities, there is no "one size fits all" solution, the difference in educational resources for agricultural security varies from county to county in the USA [29]. It is no stretch of the imagination to see that disparities exist country to country for agriculture, water supply, and food supply chains. These sectors are critical everywhere around the world, but the resources for cyberbiosecurity are not equally distributed, so a solution needs to be general and easy to implement and maintain. Authors in Duncan et al. [88], by focusing on the US food supply chain, stated that "this gap in education and training increases risks to the domestic [U.S.] food supply chain and the ultimate mission of securing the U.S. and global food supply."

Lack of understanding is a significant risk for any cyberbiosecurity system, but especially for small farms where available knowledge and resources are less than large infrastructure organizations (e.g., utility companies, and industrial farms). More needs to be done to explain cyberbiosecurity as a concept and raise awareness of the vulnerabilities it creates. Richardson et al. [2] point out that as agricultural becomes more reliant cyber-enabled systems the security of these systems is "unclear from a cyberbiosecurity perspective." This is at the same time that technology is increasingly incorporated into water supply and food supply systems, creating similar vulnerabilities [9, 34, 43, 48, 89–91]. Although, Reed and Dunaway [40] were optimistic that technology would bring solutions without any vulnerabilities.

As the size of an organization increases (e.g., industrial farms, utility water supplies, and the bioeconomy) so does complexity and difficulty in understanding how the system operates. Lack of understanding of minute details and interconnectedness are a vulnerability, as even changes to mitigate external threats can lead to unintended consequences [39]. Imagine updating security software and a bug

prevents water tanks in a system form relaying fill levels to the central control. More effort needs to be placed on understanding how the system actually operates and how best to explain that operation to the people it matters most.

This approach needs to be done on a case by case basis, as the variability in each individual systems differs. Germano [47] and Clark et al. [48] both point out that differences among organizations and utilities in the water and wastewater sectors include size (employee count and water processed), management, available resources, regulatory oversight, and even security protocols. These differences make a unified approach to cyberbiosecurity in the water sector unfeasible, as each organization or utility needs to build their own approach to match their unique operation and threats. The water distribution system for a large city is going to vary in size, available resources, and security measures from that of a small rural county. This disparity exists in the other sectors of the cyberbiosecurity as well, no two farms, food supply chains, or any other large-scale infrastructure are going to be the same as the issues each one deals with greatly varies. Understanding the needs and shortcomings of each system is critical for cyberbiosecurity.

Awareness of threats and how cyberbiosecurity systems operates is a form of threat mitigation, and several papers make the case for simply making people aware of the risks [26, 44, 45, 47, 92, 93]. Even something as simple as "understanding the threat and establishing cybersecurity governance protocols" is all it can take to protect these systems [47]. That said, understanding these complex systems is no trivial tasks. Both cyberbiosecurity and AI can benefit from the explainability assurance to make them human intelligible. Explainable AI systems are easier to understand how they operate and therefore understand what might negatively impact the system cyberbiosecurity systems, on the other hand, could be explained via machine learning techniques like clustering or even learning a Directed Acyclic Graph (DAG) of the data like Lin et al. [94] did for the SWaT dataset.

The next step for building understanding of cyberbiosecurity systems is through education and training. Richardson et al. [87] call for a standardization of the training process, in the same manner as biosafety and cybersecurity, through credentialing. They also called for integrating training into existing programs or relying on existing programs, as did [29], while others merely made a call for increasing education and awareness [95]. Another theme that emerged in the literature was a need for training across sectors in the water and agricultural industries, so employers training employees [45, 47], cross-sector training [80, 96, 97], government or university curated resources and training, both formal and informal [48, 88, 97], and even war-gaming [98].

5.4 Secure AI

Undoubtedly, one of the most important factors in ensuring the security of water distribution systems is to detect anomalies that may occur in these systems or malicious attacks that may come from adversaries. Water treatment and distribution

systems have been increasingly targeted by cyber-physical attacks in recent years [99]. This is partially due to the expansion of the Internet of Things (IoT) and proliferation of AI increasing the digitization of the decision-making processes and creating an adversarial attack opportunity following recent development in the machine learning field, which led to black-box adversarial methods that work well even with limited information [100].

The Kemuri Water Company (KWC) [101] attack in 2016 is a very important example of the risk these national infrastructures are under. The attack has resulted in more than 2.5 million records stolen, but more importantly, the attackers were able to change control data to manipulate the water supplied to the area. The attacks were halted before any public health damage occurred, nonetheless, it showed how vulnerable these infrastructures are and how important it is to ensure their safety.

Another recent, important incident was the Florida Water Supply hack in 2021 [21]. In this malicious attack, the hacker was able to gain remote access to the PLC (Programmable Logic Controller) unit that controls the sodium hydroxide level (also known as lye) of the water supplied to more than 15,000 residents in Tampa, Florida. The hacker was able to increase the amount of sodium hydroxide content of the water by 110 fold. Fortunately, the attack was mitigated before the poisonous levels of chemical diffused into the distribution network.

Both of these incidents show how important it is to detect any anomaly or malicious attacks early to mitigate, or hopefully prevent, any damage. Taormina et al. [35] investigated the vulnerabilities of these critical infrastructures in-depth in their research.

Pasqualetti et al. [102] investigated the detection and identification of CPS attacks from two different perspectives in their 2013 paper. They categorized the monitoring limitations from "graph-theoretic" and "system-theoretic" while proposing a mathematical framework for the problem's solution. The framework they proposed considers the CPS as a linear time-invariant descriptor system. They then defined a comprehensive set of assumptions and equation systems to measure and detect the corrupted signals in the system. They have also made a theoretical quantification of the limitations of both monitoring approaches to determine undetectable and unidentifiable attacks boundaries. Their paper is also one of the earliest attempts to formally describe the attack detection against CPSs and in this sense, its importance in the field is substantial.

Machine learning is a powerful and important tool for ensuring cyber-physical security. It is not surprising to see deep learning, more specifically Long-Short Term Memory Recurrent Neural Networks (LSTM-RNN), as efficient solutions to a problem with a time-dependent and high sequential relations such as attack detection [103]. Goh et al. [104] used the SWaT dataset [105] as a small-scale representation of a water treatment plant to detect anomalies and identify the sensors affected by this anomaly. They proposed to use the Cumulative Sum (CUSUM) method to mitigate the effects of an extremely unbalanced distribution of positive and negative classes (millions of negative samples to only thousands of positive samples with a sequential dataset). The SWaT dataset is a comprehensive and very

important dataset for cyber-physical security research and the contributions of the authors and supporting organizations to the field should not be left unacknowledged.

Inoue et al. [106] applied another deep learning approach in their 2017 paper. The authors used a Deep Neural Network (DNN) to evaluate the Support Vector Machine (SVM) method's performance for anomaly detection problems. The paper also made a side-by-side comparison of the two models while discussing their advantages and disadvantages. Unlike Goh et al. [104], the authors did not address the data imbalance in the paper. The researchers used the SWaT dataset and the simulation to test the models.

BATADAL is a planning and management competition for Water Infrastructures and it takes place as part of the Water Distribution Systems Analysis Symposium. This competition presents an imaginary C-Town as a water distribution network dataset to detect the real-life size and real-time, simulated data from this town (SCADA) [107]. The paper includes seven well-performing solutions to the problem on this dataset from the competitors. Others (Aghashahi et al. [108]) used a two-stage approach to solve the anomaly detection problem. In the first stage they make a feature extraction, and in the second stage they use a supervised classification method, Random Forests, to detect attack instances.

Brentan et al. [109] proposed a statistical approach to the problem. They used the sectioned nature of the problem environment and trained Recurrent Neural Networks (RNNs) to learn each district's normal behaviors and then calculated the deviation from these expected normals to measure the anomaly levels on the system.

Chandy et al. [110] used a similar two-staged approach to Aghashahi et al. [108]. Chandy et al. [110], however, first make a detection of the anomaly and then confirm or reject this detection is with a second model, a Convolutional Neural Network (CNN) Auto-Encoder, by calculating reconstruction probabilities.

Giacomoni et al. [111] proposed another two-stage approach. In the first stage, the authors created a set of rules and calculated the integrity of the rules for each instance. In the second stage, they analyzed the dataset to calculate certain thresholds of normalcy. They also proposed using Principal Component Analysis (PCA) and convex optimization routine to perform this analysis [112].

Abokifa et al. [113] proposed a three-stage model and they classified different types of attacks on each stage of the process. In the first stage, the authors used statistical methods to detect local outlier events. In the second stage, they introduced a neural network to the process to detect operational outliers. In the third stage, they focused on the global scope to detect events that might affect more than one aspect of the system with PCA.

Pasha et al. [114] introduced another three-stage method for anomaly detection. The first stage checked the consistency of the underlying rules of the water distribution system. The second stage checked each component for behavioral patterns to see if the system is following the normal patterns it is supposed to do. If any anomaly is detected in the first two stages, the third stage confirms the detections by comparing the estimations of the system made by the method.

Housh and Ohar [25] used EPANET to create a simulation of a water distribution system's behavior to calculate the difference between the SCADA and the expected

values from the simulation to detect and locate anomalies in the systems. Housh and Ohar [115] also used a similar approach to detect contamination attacks against water distribution systems with successful results.

Taormina et al. [107] have comparatively investigated all these proposed approaches, and many more, are discussed along with the advantages and disadvantages of the models. Even though the methods are very diverse, one common factor should not be unnoticed: each of the major competitors followed a direction of first discovering underlying behavioral principles of the system in some manner and then proposed ways to measure the diversion from these principles in anomalous scenarios.

The BATADAL competition provides immense contributions to the cyber-physical security field by providing a great dataset to the researchers as well as creating a valuable comparative environment for all the approaches to provide assurances methods for cyber-physical security [107], an approach (competitions) that proved successful in other areas of AI. Kravchik and Shabtai [116] investigated the attack detection problem from an ICS perspective in their 2018 paper. They used the SWaT dataset to train CNN and Long-Short Term Memory (LSTM) models to compare their effectiveness to detect anomalies. The experimental results showed that 1D CNNs can outperform RNN and LSTMs in more complex multivariate tasks.

Umer et al. [117] investigated attack detection from a distributed system. In their work, they separated the endeavor into two categories: "design-centric" and "data-centric," while proposing a model for each category. The research used the SWaT dataset [105] as a small-scale representation of a water treatment plant. The methods they proposed utilize Association Rule Mining (ARM). They also compare the advantages and disadvantages of the two approaches proposed in the paper.

Junejo and Goh [118] proposed a behavior-based machine learning approach for the detection and classification of cyber-physical attacks. Their approach promised a low false-positive rate, which some of the other approaches discussed earlier suffer from, and still provided high recall and precision. They used the SWaT dataset to evaluate the effectiveness of nine different algorithms from supervised machine learning literature ranging from Bayesian networks, naive Bayes, logistic regression, neural networks, SVM, and more while making comparisons between models for advantages and disadvantages.

Adepu and Mathur [119] proposed a Single-Stage Multi-Point (SSMP) type of attack with a distributed detection method. Even though they focus on single-stage attacks in their paper, the authors noted that they found it more effective to detect this kind of attack using the information from neighboring stages. The researchers used the SCADA dataset to create two invariants: State-Dependent (SD) and State-Agnostic (SA). Later the authors combined both invariants to create a more efficient tool for distributed detection problems.

In another paper, by Adepu and Mathur [120] authors used the SWaT dataset to investigate ways to improve cyber-physical security and attack detection problems by asking the following questions: "What attacker and attack models should be used to understand the behavior of a CPS?", "How do cyber-attacks impact a specific CPS

with respect to the number of actuators affected, state of a CPS when the attack is launched, and duration of the attack?", and "Given the response of a CPS to one or more cyber-attacks, how does one design attack detection mechanisms using the physical properties of the system?". While trying to answer these questions with experimental results the authors disclaimed the generalizability of their findings and stated that this research only targets the SWaT testbed.

This disclaimer shows a very important direction that requires more attention in the field, which is the generalization of the proposed methods since almost all of the methods we discussed so far require prior knowledge of the attack samples to be effective in the first place. The need for generalizability of the proposed approach is the utmost importance since solutions cannot wait until the attacks happen on the real systems to collect the necessary data to train the models.

Adepu and Mathur [121] must have seen this problem as well, as they tried to address it in their next work with a case study of their earlier distributed attack detection proposal [119]. Adepu and Mathur replicated real-life scenarios to test their improved attack detection mechanism and shared their findings with the strength and weaknesses of the model with an in-depth discussion [121].

As we pointed out earlier, fast adoption of automation and networking technology does not come without drawbacks. Al-Abassi et al. [122] tried to remark these issues and address the vulnerabilities created by another attack detection method while promising generalizability on the way. The researchers propose a combined model of DNN and Decision Tree with results that outperformed most of the conventional machine learning models including DNN and Random Forest. The authors also addressed the imbalanced class distribution and effective performance of the proposed approach with experimental results.

5.5 Trustworthy AI

AI is used in an increasing number of different systems, for example, autonomous vehicles, search engines, recommendation systems, medical imaging [123], public health [124], and others. It appears well-developed, yet there are still a lot of issues that need to be addressed and discussed, especially when it comes to the question can AI be trusted in "these scenarios that have life-critical consequences?" [125]. The foundation of societies, economies, and sustainable development is based on trust. If there is no trust the whole societal system would not grow or be stable [126, 127], and the same applies to cyberbiosecurity applications. Inderwildi et al. [128] discussed the impact of intelligent CPSs in energy provision and gave policy recommendations to lower potential risks. The same applies to AI systems, the idea of trustworthy AI is to build trust between users, developers, and the system itself [129].

Trust is a concept that is difficult to build, and trust in AI is even harder to address. The "black-box" characteristic is one of the most important reasons of mistrusting AI [130]. It is hard to build trust without knowing why the system

makes its decision. We need to be able to explain the results, and this leads to the importance of explainable AI (see Sect. 5.3). Another situation where trust in AI faces scrutiny is ethical decisions, such as the trolley problem. What is the priority that the system should follow? Are there any guidelines to follow? There are so many different questions to address in order to build trust.

In recent years, a significant amount of research on trustworthy AI has been conducted in different academic and industry areas (see Fig. 3). Each study focused on different aspects of trustworthy AI, for example, [131] focused on government guidelines, which advise how to establish a trustworthy AI system through rules and regulations, and other studies focused on the computational aspect of achieving trustworthy AI [132–137]. Most of the research agrees that trustworthy AI systems should include a set of properties: reliability, safety, security, privacy, availability, usability and can be extended to the following dimensions: accuracy, robustness, fairness, accountability, transparency, interpretability/explainability, and ethics [56, 125, 126, 129, 131–133, 138–141].

Trust is a complicated concept that combines numerous factors, and different researchers from various backgrounds would also see trustworthy AI from a diverging perspectives. Liu et al. [132] defined trustworthy AI from three perspectives: technical, user, and social. The system should focus on accuracy, robustness, and explainability from a technical perspective; while it should focus on availability, usability, safety, privacy, and autonomy from the user's perspective. Whereas from the social perspective, there should be a guideline or regulation regarding legality, ethics, fairness, accountability, and environmental-friendliness. To have more clear guidelines for accomplishing trustworthy AI, the EU established the High-Level Expert Group (HLEG) to provide ethical guidelines, not just principles to follow but also concrete operational steps that allow an AI developer to examine when building and deploying an AI system [131]. Zicari et al. [58] proposed a state-of-the-art process to evaluate the trustworthy AI based on applied ethics called "Z-Inspection," which is also first process in practice that HLEG defined to evaluate the trustworthiness of AI. Z-Inspection consists of three processes: set-up, access, and resolve, and each phase breaks down into different aspects to examine whether the AI systems are trustworthy.

Toreini et al. [133] pointed out that there are various AI policy frameworks to follow from different nations and organizations, and categorize those objectives into eight qualities: privacy, accountability, safety & security, transparency & explainability, fairness & nondiscrimination, human control of technology, professional responsibility and promotion of human values. They further mapped these eight qualities with four principles, including fairness, explainability, auditability, and safety. The authors separate two main technologies of trustworthiness: Data-Centric Trustworthiness and Model-Centric Trustworthiness.

Liu et al. [132] stated "Trustworthy AI are programs and systems built to solve problems like a human, which bring benefits and convenience to people with no threat or risk of harm." They focused on six dimensions in achieving trustworthy AI including safety & robustness, nondiscrimination & fairness, explainability, privacy, accountability & auditability, and environmental well-being. Instead of focusing on

policy framework or guidelines, they worked on specific computational solutions for each dimension for realizing trustworthy AI.

Li et al. [138] mentioned AI practitioners, including researchers and developers, should focus on pursuing system performance as the main goal, whereas this is not sufficient to reflect the trustworthiness of an AI system. Therefore, they proposed a methodology that takes the entire lifecycle of AI systems into consideration, from data management to model development, deployment, and all the way to monitoring and governance. For the future research direction, while adopting this systematic approach, there are side-effects due to increased learning time and slowed development by using this new approach.

We mentioned that the trustworthiness of AI is essential when it comes to AI systems related to life-critical consequences. There were incidences where critical CPSs came under attack [142] and affected the overall trust in CPSs. For example, an attack happened on a water treatment plant in Florida in 2021 and the level of sodium hydroxide in the water supply was increased over 100 times higher than usual [143]. There were also numerous cyber-attacks on Israel's water system in 2020 [144]. That exposes how vulnerable those CPSs are and the importance of the security of those systems [145–155]. There has been no lack of related research done in the area of anomaly detection in water system or its security challenges using machine learning methods [33, 107, 116, 156–190], statistical methods [191–198], or other tangential methods [106, 199–213].

Wang et al. [214] applied probabilistic model learning to probabilistically validate a real-world CPS. MR and Mathur [215] proposed "AICrit" to effectively detect anomalies in real-time with low false alarms. Another factor contributing to the complication of evaluating trustworthiness is that most of the research or review that discusses how to achieve trustworthy AI focuses more on social science topics, such as ethics and policy [59, 139]. Most of the frameworks or guidelines they proposed, however, do focus on the human factor. Uslu et al. [216] proposed a decision-making framework to manage Food-Energy-Water (FEW) resources. While developing the optimal solutions under different scenarios, they included humans in the framework to make the solutions more trustworthy. They introduced two new metrics, trust sensitivity and trust pressure, in the framework and used a game-theoretical tool to explore the relationship between trust sensitivity and the distance of community-desirable solutions.

6 Discussion

6.1 Attack Detection Models for Water Systems

Cyberbiosecurity attack/anomaly detection research in the literature mainly focused on three datasets SWaT, WADI, and BATADAL which have been introduced in Sect. 4. These three datasets have become field leading benchmarks. As a part

of the survey, we have created tables for each dataset. In order to make a fair comparison, we have used the most commonly reported statistical metrics to rank models proposed by researchers for attack/anomaly detection problem. For SWaT (Table 1) and WADI (Table 2) datasets it was F-Score (also known as F-measure, more specifically F_1 score) and for BATADAL (Table 3) we have used S score defined by Aghashahi et al. [108] and listed S_{TTD} (Time Taken for Detection) as well. For each dataset, state of the art over the years has been marked with bold fonts on Tables 1, 2, 3.

Table 1 SWaT F1-Scores[a]

Authors	Model	F1-Score	Year
Ayas and Ayas [63]	**Modified DenseNet**	**0.9999**	**2020**
Alqurashi et al. [184]	MLP	0.9900	2021
Krithivasan et al. [217]	EPCA-HG-CNN	0.9805	2020
Xu et al. [218]	ATTAIN	0.9759	2021
Li et al. [219]	**MAD-GAN**	**0.9517**	**2019**
Kravchik and Shabtai [116]	**1D CNN**	**0.9200**	**2018**
Abdelaty et al. [220]	DAICS	0.8890	2021
Elnour et al. [221]	DIF	0.8820	2020
Kravchik and Shabtai [222]	AE Frequency	0.8730	2019
Kravchik and Shabtai [116]	1D CNN	0.8710	2018
Sapkota et al. [163]	CNN + LSTM w/ WT	0.8610	2020
Perales Gómez et al. [223]	MADICS	0.8510	2020
Lin et al. [94]	TABOR	0.8230	2018
Zizzo et al. [224]	LSTM	0.8170	2019
Shalyga et al. [225]	MLP	0.8120	2018
Li et al. [226]	GAN	0.8100	2019
Shalyga et al. [225]	CNN	0.8080	2018
Inoue et al. [106]	DNN	0.8030	2017
Faber et al. [227]	CNN 1D[b]	0.8000	2021
Inoue et al. [106]	**One-class SVM**	**0.7960**	**2017**
Shalyga et al. [225]	RNN	0.7960	2018
Inoue et al. [106]	SVM	0.7960	2017
Faber et al. [227]	USAD	0.7900	2021
Faber et al. [227]	CNN 1D	0.7800	2021
Goh et al. [104]	LSTM-CUSUM	0.7754	2017
Chakraborty et al. [228]	Random Forest	0.7700	2021
Li et al. [229]	GAN-AD	0.7500	2018
Toe et al. [70]	MARS	0.7480	2020
Faber et al. [227]	LSTM-VAE	0.7200	2021
Shalyga et al. [225]	RNN	0.6900	2018
Sapkota et al. [163]	CNN	0.6500	2020

[a] *Disclaimer*: These results are not validated as a part of this research

Table 2 WADI F1-Scores[a]

Authors	Model	F1-Score	Year
Xu et al. [218]	**ATTAIN**	**0.7444**	**2021**
Goh et al. [104]	**LSTM-CUSUM**	**0.6595**	**2017**
Li et al. [219]	MAD-GAN	0.5945	2019
Faber et al. [227]	CNN 1D	0.5400	2021
Faber et al. [227]	CNN 1D	0.5200	2021
Faber et al. [227]	USAD	0.4300	2021
Faber et al. [227]	LSTM-VAE	0.2800	2021

[a] *Disclaimer*: These results are not validated as a part of this research

Table 3 BATADAL S Scores[a]

Authors	Model	S Score	S_{TTD} Score	Year
Brentan et al. [230]	**Statistical analysis**	**0.9730**	**0.1900**	**2021**
Housh and Ohar [25]	**MILP**	**0.9700**	**0.9650**	**2018**
Abokifa et al. [160]	ANN and PCA	0.9660	0.9840	2019
Abokifa et al. [113]	**ANN**	**0.9490**	**0.9580**	**2017**
Ramotsoela et al. [231]	QDA	0.9400	0.9500	2019
Tsiami and Makropoulos [232]	TGCN	0.9310	0.9340	2021
Giacomoni et al. [111]	PCA	0.9270	0.9360	2017
Ramotsoela et al. [231]	MD	0.9100	0.9000	2019
Ramotsoela et al. [231]	iForest	0.9000	0.8600	2019
Brentan et al. [109]	RNN	0.8940	0.8570	2017
Ramotsoela et al. [231]	LOF	0.8700	0.8500	2019
Ramotsoela et al. [231]	SOD	0.8600	0.8300	2019
Mahmoud et al. [233]	SVM	0.8200	0.8400	2022
Mahmoud et al. [233]	3NN	0.8200	0.7500	2022
Mahmoud et al. [233]	RForest	0.8200	0.7800	2022
Mahmoud et al. [233]	XGBoost	0.8200	0.7500	2022
Mahmoud et al. [233]	BOSS	0.8200	0.7100	2022
Chandy et al. [110]	Convolutional variational auto-encoder	0.8000	0.8300	2017
Gjorgiev and Gievska [193]	VAE-D	0.8000	0.9750	2020
Gjorgiev and Gievska [193]	VAE-D-C	0.7780	0.9870	2020
Gjorgiev and Gievska [193]	LSTM-VAE-C	0.7780	0.9990	2020
Pasha et al. [114]	Statistical analysis	0.7730	0.8850	2017
Gjorgiev and Gievska [193]	LSTM-VAE-2E-C	0.7610	1.0000	2020

(continued)

Table 3 (continued)

Authors	Model	S Score	S_{TTD} Score	Year
Mahmoud et al. [233]	5NN	0.7600	0.6430	2022
Choi et al. [234]	SVM	0.7540	0.7220	2020
Gjorgiev and Gievska [193]	VAE-ReEncoder	0.7520	0.9350	2020
Mahmoud et al. [233]	7NN	0.7500	0.6345	2022
Ramotsoela et al. [231]	Naive Bayes	0.7500	1.0000	2019
Choi et al. [234]	ANN	0.7490	0.7590	2020
Gjorgiev and Gievska [193]	LSTM-VAE	0.7350	0.9790	2020
Mahmoud et al. [233]	1NN	0.7300	0.5720	2022
Gjorgiev and Gievska [193]	VAE-ReEncoder-C	0.7260	0.9400	2020
Gjorgiev and Gievska [193]	CNN-VAE-C	0.7130	0.9310	2020
Ramotsoela et al. [231]	OSVM	0.7100	0.6900	2019
Ramotsoela et al. [231]	LDA	0.6700	0.6500	2019
Gjorgiev and Gievska [193]	LSTM-VAE-2E	0.6640	0.8200	2020
Choi et al. [234]	ELM	0.5910	0.9410	2020
Aghashahi et al. [108]	RForest	0.5340	0.4290	2017
Gjorgiev and Gievska [193]	CNN-VAE	0.5230	0.5430	2020
Choi et al. [234]	5NN	0.4180	0.3230	2020

[a] *Disclaimer:* These results are not validated as a part of this research

Throughout the years efficiency of the neural network based models have drastically increased over numerous problems and attack/detection is one of them as well. Looking at highest ranked models on the SWaT F1-Scores Table 1, it can be seen that deep learning had a huge impact on the problem and following the success of Inoue et al. [106] with One-class SVM, in last 4 years breakthroughs were achieved using Deep Learning models Kravchik and Shabtai [116], Li et al. [219] and Ayas and Ayas [63]. This dominance can further be verified with the successful state of the art models developed by Goh et al. [104] and Xu et al. [218], once again using DNN models.

When it comes to the BATADAL dataset the picture slightly changes. Neural Network based models are still very effective on solving attack detection problem with BATADAL as well but they are not as dominant as they are with the other two datasets. Various types of approaches to the problem from many researchers provide a great understanding of the chaotic nature of data-driven problems on large physical systems. Dynamical essence of these systems requires researchers to approach the problem from many angles to ensure the models they would create to be trustworthy and secure. Some of the most successful researches to achieve these feats were, Abokifa et al. [113], Housh and Ohar [25] and Brentan et al. [230] as the state of the art holders.

6.2 *Assessing the Cyberbiosecurity Literature*

In this section, we discuss cyberbiosecurity further because it is a new discipline and there are different takes on exactly what it is. Unfortunately, most of the literature writes about cyberbiosecurity in a manner similar to cybersecurity for biological applications [8, 39, 81, 84, 90, 92, 95, 235–239].

This is not a fault, the focus of cyberbiosecurity is biology or related applications; however, most of the literature does not adequately define what sets cyberbiosecurity apart from IT or Computer Science in the life sciences. Gillum et al. [97] expressed a similar concern with the issues in the term "biosecurity," established fourteen years prior to their work. Multiple papers in the literature call for action or collaboration—"We call for analyses and publications to fully scope cyberbiosecurity and identify a comprehensive strategy to establish the discipline's goals and objectives" [2] and others, as called out by Drape et al. [29] and seen in Murch and DiEuliis [26]. This call from Richardson et al. [2] makes it seem like the field is still in the early planning stages, but this is not entirely true as there are papers that focus on concrete examples, lie case studies, surveys, and even one where the authors initiated an attack on a synthetic DNA supply chain that went undetected [29, 80, 86, 93, 97, 238].

Cyberbiosecurity systems are rooted in the physical sciences, but they can include pure information systems like databases for pathogens, genomics data, and land use data [4, 44, 83, 235]. We focus, however, on the physical supply chains and infrastructure, specifically water and food supply systems. Here, cyberbiosecurity secures supply through "the design of digital strategies, business models, technologies, standards and regulations" [240]. This does not exclude systems that rely on data, as even food systems depend on sharing and gathering insights from data. For example, in Duncan et al. [80] the authors discuss the need for sharing and protecting data to "design promising agricultural and food systems to better meet consumers' need." Data is just as much a part of physical systems.

Water systems are open to both natural anomalies and intentional attacks, something highlighted by Schmale III et al. [23], in their paper on a water supply system that is subject to harmful algal blooms, remote monitoring and control are incorporated to help ensure the water stays safe for drinking. However, this opens the system up to cyber-attacks, so cyberbiosecurity measures need to be taken to monitor and mitigate both sources of issues to ensure the safety of the water.

These systems are complex and multifaceted, which makes protections harder to implement and formalize, and this sentiment is highlighted in Duncan et al. [9] where the authors state current protections are not enough and "do not broadly exist across the food and agricultural system," and the "conversation on cyber security on the U.S. food and agricultural system (cyberbiosecurity) is incomplete and disjointed." There is a critical need to better incorporate cyberbiosecurity into the water and food supply chain infrastructures. Something easier said than done as these systems have multiple layers of weaknesses at the software level, the interface of cyber and physical, and the biological level. A sentiment that was

expressed in Farbiash and Puzis [238] for the synthetic DNA supply chain, as those authors demonstrated an attack can bypass cybersecurity and biosecurity screenings to generate an attack based on gene editing in the synthetic data. In Bernal et al. [28], the work presented used bacteria in a DDoS style attack to demonstrate the unique risks to cyberbiosecurity that traditional cybersecurity measures cannot accommodate. These papers highlight the fact that there are biological exploits available to cyberbiosecurity systems an attacker can use without ever having physical access to a system. The multifaceted supply chains allow for multifaceted attacks that can slip through the cracks of traditional cybersecurity and biosecurity efforts.

6.3 Adoption of AI Assurance for Cyberbiosecurity

The goal of AI assurance is to mitigate any potential drawbacks or failures of AI in high-stakes applications. Assurance is a way of validating AI operates in a human-centered manner, and likewise the goals of cyberbiosecurity are to protect people from biological threats in many forms, they just happen to focus on cyber-systems and CPSs specifically. Despite this alignment of goals, we see little direct connections between cyberbiosecurity and AI in the surveyed papers (see the separation of cyberbiosecurity from the other papers in Fig. 4). There are, however, a handful of cyberbiosecurity papers we found that do overlap in topic with AI assurance, even if there is no connection via citations. Most of these papers deal with trustworthiness and safety [8, 28, 84, 241], and in fact these are also the most common assurances in the literature (see Fig. 3). Two of these papers also focus on fairness [84, 241], a little more surprising because fair AI was the least common assurance we found (again, see Fig. 3). There is one paper that focuses on explainability, specifically data and model transparency, in cyberbiosecurity [44], and how explainability ties more to security. The last paper focuses solely on trustworthiness in cyberbiosecurity [242].

Safety is a key AI assurance pillar (see Sect. 5.2), followed closely by trustworthiness (Sect. 5.5), that applies to cyberbiosecurity. The efforts of all the others are done in order to ensure the safety of the system or in the trust that the system operates in a safe manner. Ethical and fair AI (Sect. 5.1) ensures the AI system makes decisions that are correct and benefit everyone impacted equally, letting users trust that the AI makes safe decisions. Explainability (Sect. 5.3) gives us understanding of how the system operates and why it makes the decisions it does, letting users trust that the AI operates as it should to ensure the safety of those impacted. Secure AI (Sect. 5.4) ensures that if problems arise (anomalies or attacks) that the AI can handle them, either by correcting or mitigating negative effects, letting users trust that the AI system negates or limits possible harm to those impacted. Everything is done so we can trust the safety of the system.

Safety in cyberbiosecurity is mostly concerned with biosafety, or the protection from biological threats. We believe there should be more focus in the literature

on food and water safety from a cyberbiosecurity perspective, especially as more technology is adopted in the water and agriculture sectors. However, there are some existing safety measures that can be adopted, like the Hazard Analysis and Critical Control Points (HACCP) for food safety and management which could be used as a starting point for safety assurances [9, 88].

Policy and regulations need to be part of the cyberbiosecurity solution, in part for the need of creating standard practices and metrics across the whole bioeconomy, and in part because cyberbiosecurity threats pose national and international security risks [243]. Cyberbiosecurity should be part of the national strategy for cybersecurity, part of the "Defend Forward" idealogy of national security [244]. This approach, however, requires the need for understanding the cyberbiosecurity field to create regulation and policy for federal agencies, something which is still lacking as "cyberbiosecurity roles, practices and metrics have not been defined and federal agencies appear uncertain regarding how to proceed" [93, 245].

The current state of the cyberbiosecurity literature focuses more on creating systems of awareness or best practices for mitigating security or safety threats, and there is little direct discussion on using explainable AI for cyberbiosecurity. Explainable AI lacks discourse in the cyberbiosecurity literature but is discussed frequently in the medical AI domain, where the goal is to create trust in AI in order to facilitate adoption by medical practitioners and to create transparency and traceability in the decisions made by the AI [246]. Explainable AI also allows for the combination of an interpretable, knowledge-based approach with that of an efficient neural based approach [247]. This means explainable AI is a way of augmenting human understanding of a problem when it uses models designed for human comprehension.

The augmentation of human intelligent via explainable AI feels like a particularly fitting application of AI for cyberbiosecurity. There is still more challenges to be addressed in the domain of explainable AI to show applicability in real-world deployments [246]; however, it does offer a lot of promise in applications where decisions are high-stakes, such as critical infrastructure including agricultural, food, and water supply chains. Richardson et al. [2] called for the implementation of "frameworks to facilitate responsible application of AI techniques to biology" and explainable AI is one way to do so.

This is particularly important to cyberbiosecurity and parts of the bioeconomy, where the sheer size and complexity of systems creates the potential for unintentional harm when trying to mitigate threats [22, 39]. Training and education of these systems (AI or otherwise) become a form of ensuring the continued safe operation of these complex systems. Training and education are also a form of creating awareness of threat mitigation to help ensure security. This is a common theme in the cyberbiosecurity literature [26, 29, 44, 45, 47, 80, 87, 88, 92, 95, 97].

All the pillars eventually boil down to ensuring trust that AI and cyberbiosecurity systems operate as intended. Section 5.5 discussed the connection of AI assurance to trustworthy AI. Society and the bioeconomy, in general, are built on trust, and if we do not trust them we will not use or participate in their activities. The same

goes for AI in cyberbiosecurity, trust needs to be built so operators and all parties involved use them.

Developments in AI for cybersecurity and cyber-physical security could protect water, food, or other supply chains from intentional interference, while developments in AI for anomaly detection could protect the supply from natural phenomena [23, 25, 94, 102, 104, 106, 114, 119, 121, 225, 248–254]. Despite a clear alignment of incentives, there is not much direct overlap between these approaches in the cyberbiosecurity literature (see the separation of between cyberbiosecurity and attack/anomaly detection in Fig. 4). We conclude that although more of the cyberbiosecurity papers clearly make a call for action [2, 26, 29, 255], there is at best merely a brief attempt over existing solutions like the National Institute of Standards and Technology (NIST) cybersecurity framework [43, 47, 95, 256]. The safety and continuing function of any and all systems in the bioeconomy are important but "currently protections are minimal and do not broadly exist across the food and agricultural system" [9].

6.4 Merging the Water Security and Cyberbiosecurity Fields

Similar to AI assurance, there is not a large direct link in the literature between cyberbiosecurity and water systems. There is one series of links from cyberbiosecurity to water systems via Mueller [22], Schmale III et al. [23], Moyer et al. [24], and Housh and Ohar [25]. When we broadened our definition of cyberbiosecurity a little more from the literature we see a broader connection of papers that link the topic with water supply systems [6, 9, 23, 47, 48, 257]. What is also interesting to note is that none of these papers uses the open-source datasets we discussed in Sect. 4, instead these papers focus on broad topics of water within the food and agriculture sector [6, 9, 257] or the security of water sources [23, 47, 48]. Most of the water supply-related papers deal with security and attack/anomaly detection, aligning them more with AI assurance, but we feel they apply just as much to cyberbiosecurity as well.

There is not much existing cyber or cyber-physical security knowledge within the cyberbiosecurity field [2, 8, 29, 45, 86–88, 97]. This makes the openness of water supply testbeds and AI research critical, as these technologies can be developed and tested open-source in view of researchers focusing on cyberbiosecurity. More emphasis of the cyberbiosecurity research should be placed on using the open-source water testbeds from Sect. 4. This is the only way that water security (as a form of cyberbiosecurity) research can be performed using relevant data, and it also allows for training and hands-on experience, something a large portion of the literature called for [26, 29, 44, 45, 47, 80, 87, 88, 92, 95, 97]. This development of human understanding of cyberbiosecurity and water systems is a form of explainability and it significantly benefits from open-source data on how these systems operate.

6.5 Recommendations and Future Direction

Much of the work regarding AI assurance and cyberbiosecurity occurred in the last few years and developed separately. Figure 4 shows one link connecting cyberbiosecurity to water systems, which is then tied to the large web of anomaly and attack detection papers. Cyberbiosecurity research, however, still has a long way to meet its goal of wider adoption, and while we cannot speak for all possible sources of cross-collaboration, the expansion of cyberbiosecurity into the domains of water supply systems and AI assurance is wide open for future research.

Continuing the thread of expanding the research outside its immediate domain, cyberbiosecurity has a lot to gain from embracing open-source water supply testbeds. For one, the domain of water security is directly applicable to cyberbiosecurity, despite not making up much of the research. The literature mostly focuses on biology applications, but this feels narrow and collaborating with the established field of water security would be a great way to apply all those lessons learned to cyberbiosecurity. Many of the papers in the cyberbiosecurity literature call for more training, education, and hands-on experience. Open-source testbeds are ideal for developing resources for training and education, as well as developing new research into secure AI and other forms of AI assurance.

The goals of assurance are to validate AI aligns with the values of users impacted by an AI system, and likewise the goal of cyberbiosecurity is to protect users and citizens impacted by a biological system. AI has been instrumental in multiple agricultural applications [258–260] and offers many solutions to the threats of cyberbiosecurity but also includes several downsides; assurance nonetheless offers a way to apply AI to maximize its benefits while mitigating potential pitfalls. AI assurance should also be broadened to focus on the entirety of the system AI is deployed in, not just the assurance of the AI itself. For example, both applying AI to ensure the safety of drinking water via water quality monitoring and applying evaluation procedures to ensure the AI is operating properly are forms of assurance. In short, the cyberbiosecurity field should adopt AI measures to meet its goals and use AI assurance to validate both the AI employed is working properly and that the larger system the AI is used in is also operating properly.

7 Conclusions

In this survey, we investigated academic papers at the intersection of AI assurance, cyberbiosecurity, water and food supply systems. We assessed the application, both current and potential, of AI assurance to problems in cyberbiosecurity, specifically focusing on water and food supply systems. The survey focused on journal articles, conference proceedings, dissertations, books and book chapters, and industry white papers published from 2000 to April 2022 and at the intersection of two or more of the mentioned sectors.

A survey landscape (Sect. 2) was performed for an overview of the literature, showing most of the papers included were published since 2016, as researchers started applying AI more broadly and investigating AI assurance. Soon after in 2017, the field of cyberbiosecurity had traction and more water supply system papers were published. The increase in water supply papers since 2016 seems in part due to the start of open-source testbeds (SWaT in 2015, WADI in 2016, BATADAL in 2016, Smod in 2017, and DHALSIM in 2020), and because we specifically focused on papers that intersected with AI and cyberbiosecurity fields, both of which have seen sharp increases in the past few years. Although, looking at Fig. 4, we see there is little connection between the literature of cyberbiosecurity with the other sectors. We discussed how the papers covering these topics connected and how AI assurances apply in these fields, followed by our recommendations for future directions.

In the previous sections, we discussed the six pillars of AI assurance [1], the importance of each pillar, and the effects of the papers surveyed on water distribution systems and their applications. Figure 3, however, shows this distribution is not uniform. The pillars of Ethical AI and Fair AI were neglected, while the importance of these aspects kept growing over the last several years. This shows a great gap and opportunity for research in Ethical and Fair AI for agricultural and water systems.

We found less collaboration among the fields of AI assurance, cyberbiosecurity, and water or food supply systems than we initially expected. Figure 4 shows this disjoint well, and the literature for cyberbiosecurity does not directly discuss AI much, let alone AI assurance. The cyberbiosecurity definition should adapt a little more, as it feels too focused on cybersecurity for the life sciences. There is some acknowledgement that the current literature is not broad enough [9], especially when there are biological processes that can be exploited [28, 238].

Further research should emphasize collaboration across sectors and the use of open-source datasets and testbeds. The call for collaboration already exists with the cyberbiosecurity field, and one of our proposed solutions to that is publishing open-source datasets online. These open the field to broader research and hands-on training and experience, both of which have been expressed as needs for the cyberbiosecurity field. There are unique challenges, though these require expertise from biology, CPSs, and other domain specific knowledge for a desired application.

Lastly, we recommend that the cyberbiosecurity field adopts AI and AI assurances practices for better security while maintaining safe and trustworthy operations of these complex biological systems. There has been a lot of prior research applying AI for cybersecurity, and this would be a natural extension to incorporate into cyberbiosecurity. AI also offers more robust monitoring and an ability to make corrective actions, but this is not without issue as AI creates new vulnerabilities or failure modes. AI assurance can help mitigate these and help ensure the proper function of the overall cyberbiosecurity system.

Acknowledgments This work was supported in part by funding from Deloitte Touche Tohmatsu Limited.

We acknowledge the Center for Advanced Innovation in Agriculture (CAIA) at Virginia Tech and the Intelligent Systems Division (ISD) at The Hume Center for National Security and Technology, both for their support.

Additionally, a word of thanks to the members of Virginia Tech's A3 Research Lab (https://ai.bse.vt.edu/) for their inputs and feedback. Lastly, this work would of not been possible without the involvement of Dr. Susan Duncan (may she rest in peace) – to whom we dedicate this work.

References

1. F.A. Batarseh, L. Freeman, C.H. Huang, A survey on artificial intelligence assurance. J. Big Data **8**(1), 1–30 (2021)
2. L.C. Richardson, N.D. Connell, S.M. Lewis, E. Pauwels, R.S. Murch, Cyberbiosecurity: a call for cooperation in a new threat landscape. Front. Bioeng. Biotechnol. **7**, 99 (2019a)
3. J. Ayling, A. Chapman, Putting AI ethics to work: are the tools fit for purpose? AI Ethics, 1–25 (2021)
4. G.B. Frisvold, S.M. Moss, A. Hodgson, M.E. Maxon, Understanding the us bioeconomy: A new definition and landscape. Sustainability **13**(4), 1627 (2021)
5. The White House, National bioeconomy blueprint, April 2012. Industrial Biotechnology **8**(3), 97–102 (2012)
6. A. Aguilar, R. Wohlgemuth, T. Twardowski, Preface to the special issue bioeconomy (2018a)
7. Engineering National Academies of Sciences, Medicine, et al., *Safeguarding the Bioeconomy* (National Academies Press, 2020)
8. K.M. Berger, Addressing cyber threats in biology. IEEE Secur Privacy **18**(3), 58–61 (2020)
9. S.E. Duncan, R. Reinhard, R.C. Williams, F. Ramsey, W. Thomason, K. Lee, N. Dudek, S. Mostaghimi, E. Colbert, R. Murch, Cyberbiosecurity: A new perspective on protecting us food and agricultural system. Front. Bioeng. Biotechnol. **7**, 63 (2019)
10. R.A. Kemmerer, Cybersecurity, in *Proceedings of the 25th International Conference on Software Engineering, 2003* (IEEE, 2003), pp. 705–715
11. J.A. Lewis, Cybersecurity and critical infrastructure protection. Center Strategic Int. Stud. **1**, 12 (2006)
12. Department of Homeland Security, A glossary of common cybersecurity terminology. national initiative for cybersecurity careers and studies: Department of homeland security. http://niccs.us-cert.gov/glossary#letter_c (2022). Accessed: 2022-02-23
13. Z. Hu, J. Shi, Y. Huang, J. Xiong, X. Bu, Ganfuzz: a gan-based industrial network protocol fuzzing framework, in *Proceedings of the 15th ACM International Conference on Computing Frontiers* (2018), pp. 138–145
14. K. Lamshöft, T. Neubert, C. Krätzer, C. Vielhauer, J. Dittmann, Information hiding in cyber physical systems: Challenges for embedding, retrieval and detection using sensor data of the swat dataset, in *Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security* (2021), pp. 113–124
15. M. Dietz, M. Vielberth, G. Pernul, Integrating digital twin security simulations in the security operations center, in *Proceedings of the 15th International Conference on Availability, Reliability and Security* (2020), pp. 1–9
16. E.A. Lee, Cyber physical systems: Design challenges, in *2008 11th IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing (ISORC)* (IEEE, 2008), pp. 363–369
17. N. Jazdi, Cyber physical systems in the context of industry 4.0, in *2014 IEEE International Conference on Automation, Quality and Testing, Robotics* (IEEE, 2014), pp. 1–4
18. J. Waage, J.D. Mumford, Agricultural biosecurity. Philos. Trans. R. Soc. B Biol. Sci. **363**(1492), 863–876 (2008)

19. FAO, Biosecurity in food and agriculture. https://www.fao.org/3/Y8453E/Y8453E.htm (2003). Accessed: 2022-02-26
20. S. Hinchliffe, J. Allen, S. Lavau, N. Bingham, S. Carter, Biosecurity and the topologies of infected life: from borderlines to borderlands. Trans. Inst. Brit. Geogr. **38**(4), 531–543 (2013)
21. J. Peiser, A hacker broke into a florida town's water supply and tried to poison it with lye, police said (2021). https://www.washingtonpost.com/nation/2021/02/09/oldsmar-water-supply-hack-florida/
22. S. Mueller, Facing the 2020 pandemic: What does cyberbiosecurity want us to know to safeguard the future? Biosafety Health **3**(01), 11–21 (2021)
23. D.G. Schmale III, A.P. Ault, W. Saad, D.T. Scott, J.A. Westrick, Perspectives on harmful algal blooms (habs) and the cyberbiosecurity of freshwater systems. Front. Bioeng. Biotechnol., 128 (2019)
24. J. Moyer, R. Dakin, R. Hewman, D. Groves, The case for cyber security in the water sector. J. Am. Water Works Assoc. **101**(12), 30–32 (2009)
25. M. Housh, Z. Ohar, Model-based approach for cyber-physical attack detection in water distribution systems. Water Research **139**, 132–143 (2018)
26. R. Murch, D. DiEuliis, Mapping the cyberbiosecurity enterprise. Front. Bioeng. Biotechnol., 235 (2019)
27. T. Dixon, The grey zone of cyber-biological security. International Affairs **97**(3), 685–702 (2021)
28. S.L. Bernal, D.P. Martins, A.H. Celdrán, Distributed denial of service cyberbioattack affecting bacteria-based biosensing systems, in *2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)* (IEEE, 2020), pp. 279–282
29. T. Drape, N. Magerkorth, A. Sen, J. Simpson, M. Seibel, R.S. Murch, S.E. Duncan, Assessing the role of cyberbiosecurity in agriculture: A case study. Front. Bioeng. Biotechnol., 742 (2021)
30. C. Perakslis, Cyberbiosecurity, ecopsychology, and beyond: Our formidable pit community [last word]. IEEE Technol. Soc. Mag. **39**(4), 84–84 (2020)
31. J. Goh, S. Adepu, K.N. Junejo, A. Mathur, A dataset to support research in the design of secure water treatment systems, in *International Conference on Critical Information Infrastructures Security* (Springer, 2016), pp. 88–99
32. T. Cruz, P. Simões, Down the rabbit hole: Fostering active learning through guided exploration of a scada cyber range. Applied Sciences **11**(20), 9509 (2021)
33. Q. Lin, S. Verwer, R. Kooij, A. Mathur, Using datasets from industrial control systems for cyber security research and education, in *International Conference on Critical Information Infrastructures Security* (Springer, 2019), pp. 122–133
34. C.M. Ahmed, V.R. Palleti, A.P. Mathur, Wadi: a water distribution testbed for research in the design of secure cyber physical systems, in *Proceedings of the 3rd International Workshop on Cyber-Physical Systems for Smart Water Networks* (2017), pp. 25–28
35. R. Taormina, S. Galelli, N.O. Tippenhauer, E. Salomons, A. Ostfeld, Characterizing cyber-physical attacks on water distribution systems. J. Water Resour. Plan. Manag. **143**(5), 04017009 (2017)
36. A. Ostfeld, E. Salomons, L. Ormsbee, J.G. Uber, C.M. Bros, P. Kalungi, R. Burd, B. Zazula-Coetzee, T. Belrain, D. Kang, et al., Battle of the water calibration networks. J. Water Resour. Plan. Manag. **138**(5), 523–532 (2012)
37. P.M. Laso, D. Brosset, J. Puentes, Dataset of anomalies and malicious acts in a cyber-physical subsystem. Data Brief **14**, 186–191 (2017)
38. A. Murillo, R. Taormina, N. Tippenhauer, S. Galelli, Co-simulating physical processes and network data for high-fidelity cyber-security experiments, in *Sixth Annual Industrial Control System Security (ICSS) Workshop* (2020), pp. 13–20
39. B.C. Wintle, C.R. Boehm, C. Rhodes, J.C. Molloy, P. Millett, L. Adam, R. Breitling, R. Carlson, R. Casagrande, M. Dando, et al., Point of view: A transatlantic perspective on 20 emerging issues in biological engineering. Elife **6**, e30247 (2017)

40. J.C. Reed, N. Dunaway, Cyberbiosecurity implications for the laboratory of the future. Front. Bioeng. Biotechnol., 182 (2019)
41. J.M. Bartoszewicz, A. Seidel, B.Y. Renard, Interpretable detection of novel human viruses from genome sequencing data. NAR Genomics Bioinforma. **3**(1), lqab004 (2021)
42. A. Salam, Internet of things for sustainability: perspectives in privacy, cybersecurity, and future trends, in *Internet of Things for Sustainable Community Development* (Springer, 2020), pp. 299–327
43. M. Walsh, W. Streilein, Security measures for safeguarding the bioeconomy. Health Security **18**(4), 313–317 (2020)
44. S.B. Jordan, S.L. Fenn, B.B. Shannon, Transparency as threat at the intersection of artificial intelligence and cyberbiosecurity. Computer **53**(10), 59–68 (2020)
45. F. Ramsey, H. Seyyedhasani, Cyber attacks in agriculture: protecting your farm and small business with cyberbiosecurity
46. L. Freeman, A. Rahman, F.A. Batarseh, Enabling artificial intelligence adoption through assurance. Social Sciences **10**(9), 322 (2021)
47. J. Germano, *Cybersecurity Risk & Responsibility in the Water Sector* (American Water Works Assn, 2018)
48. R.M. Clark, S. Panguluri, T.D. Nelson, R.P. Wyman, Protecting drinking water utilities from cyberthreats. J. Am. Water Works Assoc. **109**(INL/JOU-16-39302) (2017)
49. A. Aguilar, R. Wohlgemuth, T. Twardowski. Perspectives on bioeconomy (2018)
50. D. Wakabayashi, Self-driving uber car kills pedestrian in Arizona, where robots roam. The New York Times **19**(03) (2018)
51. A. Wilk, Teaching AI, ethics, law and policy (2019)
52. C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intelli. **1**(5), 206–215 (2019)
53. C. Rudin, C. Wang, B. Coker, The age of secrecy and unfairness in recidivism prediction. Preprint (2018). arXiv:1811.00731
54. J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias, in *Ethics of Data and Analytics* (Auerbach Publications, 2016), pp. 254–264
55. L.K.J.A. J. Larson, S. Mattu, How we analyzed the compas recidivism algorithm. ProPublica (2016)
56. M. Arnold, R.K. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilović, R. Nair, K.N. Ramamurthy, A. Olteanu, D. Piorkowski, et al., Factsheets: Increasing trust in AI services through supplier's declarations of conformity. IBM J. Res. Dev. **63**(4/5), 6–1 (2019)
57. P. Laplante, D. Milojicic, S. Serebryakov, D. Bennett, Artificial intelligence and critical systems: from hype to reality. Computer **53**(11), 45–52 (2020)
58. R.V. Zicari, J. Brodersen, J. Brusseau, B. Düdder, T. Eichhorn, T. Ivanov, G. Kararigas, P. Kringen, M. McCullough, F. Möslein, et al., Z-inspection®: a process to assess trustworthy AI. IEEE Trans. Technol. Soc. **2**(2), 83–97 (2021)
59. C. Grady, S. Rajtmajer, L. Dennis, When smart systems fail: the ethics of cyber-physical critical infrastructure risk. IEEE Trans. Technol. Soc., 6–14 (2021)
60. R.A. Calvo, D. Peters, S. Cave, Advancing impact assessment for intelligent systems. Nature Mach. Intell. **2**(2), 89–91 (2020)
61. C.M. Hudson, N.D. Pattengale, R.K. Iyer, Z.T. Kalbarczyk, N. Alli, Genomic and synthetic biology digital biosecurity, in *Pacific Symposium On Biocomputing 2022* (World Scientific, 2021), pp. 402–406
62. M. Gardezi, R. Stock, Growing algorithmic governmentality: Interrogating the social construction of trust in precision agriculture. J. Rural Stud. **84**, 1–11 (2021)
63. S. Ayas, M.S. Ayas, A modified densenet approach with nearmiss for anomaly detection in industrial control systems. Multimedia Tools Appl., 1–14 (2021)
64. C. Rodríguez Martínez, M. Quiñones-Grueiro, C. Verde, O. Llanes-Santiago, A novel approach for detection and location of cyber-attacks in water distribution networks, in *International Workshop on Artificial Intelligence and Pattern Recognition* (Springer, 2021), pp. 79–90

65. Y. Wu, S. Liu, A review of data-driven approaches for burst detection in water distribution systems. Urban Water J. **14**(9), 972–983 (2017)
66. H.H. Addeen, Y. Xiao, J. Li, M. Guizani, A survey of cyber-physical attacks and detection methods in smart water distribution systems. IEEE Access **9**, 99905–99921 (2021)
67. N. Tuptuk, P. Hazell, J. Watson, S. Hailes, A systematic review of the state of cyber-security in water systems. Water **13**(1), 81 (2021)
68. S. Athalye, C.M. Ahmed, J. Zhou, A tale of two testbeds: a comparative study of attack detection techniques in cps, in *International Conference on Critical Information Infrastructures Security* (Springer, 2020), pp. 17–30
69. M. Abdelaty, R. Doriguzzi-Corin, D. Siracusa, Aads: A noise-robust anomaly detection framework for industrial control systems, in *International Conference on Information and Communications Security* (Springer, 2019), pp. 53–70
70. T.T. Toe, L.H. Yi, E.F.M. Josephlal, Advanced predictive techniques for detection of cyber-attacks in water infrastructures, in *2020 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)* (IEEE, 2020), pp. 1–6
71. S. Abba, V. Nourani, G. Elkiran, Multi-parametric modeling of water treatment plant using ai-based non-linear ensemble. J. Water Supply Re. Technol. Aqua **68**(7), 547–561 (2019)
72. M. Al-Yaari, T.H. Aldhyani, S. Rushd, Prediction of arsenic removal from contaminated water using artificial neural network model. Applied Sciences **12**(3), 999 (2022)
73. A. Jain, L.E. Ormsbee, Short-term water demand forecast modeling techniques–conventional methods versus AI. J. Am. Water Works Assoc. **94**(7), 64–72 (2002)
74. L. Karamoutsou, A. Psilovikos, Deep learning in water resources management: The case study of kastoria lake in greece. Water **13**(23), 3364 (2021)
75. L. Nishi, M. Baesso, R. Santana, P. Fregadolli, D. Falavigna, A. Falavigna-Guilherme, Investigation of cryptosporidium spp. and giardia spp. in a public water-treatment system. Zoonoses Public Health **56**(5), 221–228 (2009)
76. M. Florjanič, J. Kristl, Microbiological quality assurance of purified water by ozonization of storage and distribution system. Drug Dev. Ind. Pharm. **32**(10), 1113–1121 (2006)
77. U. Gentile, S. Marrone, F. De Paola, R. Nardone, N. Mazzocca, M. Giugni, Model-based water quality assurance in ground and surface provisioning systems, in *2015 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC)* (IEEE, 2015), pp. 527–532
78. D. Ghernaout, B. Ghernaout, On the concept of the future drinking water treatment plant: algae harvesting from the algal biomass for biodiesel production—a review. Desalin. Water Treat. **49**(1-3), 1–18 (2012)
79. I. Montalvo, J. Izquierdo, R. Pérez, M.M. Tung, Particle swarm optimization applied to the design of water supply systems. Comput. Math. Appl. **56**(3), 769–776 (2008)
80. S.E. Duncan, B. Zhang, W. Thomason, M. Ellis, N. Meng, M. Stamper, R. Carneiro, T. Drape, Securing data in life sciences—a plant food (edamame) systems case study. Front. Sustain., 10 (2020)
81. A. Adler, J. Beal, M. Lancaster, D. Wyschogrod, Cyberbiosecurity and public health in the age of covid-19, in *Emerging Threats of Synthetic Biology and Biotechnology* (Springer, Dordrecht, 2021), pp. 103–115
82. D. Greenbaum, Cyberbiosecurity: An emerging field that has ethical implications for clinical neuroscience. Camb. Q. Healthc. Ethics **30**(4), 662–668 (2021)
83. J. Caswell, J.D. Gans, N. Generous, C.M. Hudson, E. Merkley, C. Johnson, C. Oehmen, K. Omberg, E. Purvine, K. Taylor, et al., Defending our public biological databases as a global critical infrastructure. Front. Bioeng. Biotechnol. **7**, 58 (2019)
84. J. Li, H. Zhao, L. Zheng, W. An, Advances in synthetic biology and biosafety governance. Front. Bioeng. Biotechnol. **9**, 173 (2021)
85. P.M. Ney, Securing the future of biotechnology: A study of emerging bio-cyber security threats to dna-information systems. Ph.D. thesis (2019)
86. K. Millett, E. Dos Santos, P.D. Millett, Cyber-biosecurity risk perceptions in the biotech sector. Front. Bioeng. Biotechnol. **7**, 136 (2019)

87. L.C. Richardson, S.M. Lewis, R.N. Burnette, Building capacity for cyberbiosecurity training. Front. Bioeng. Biotechnol. **7**, 112 (2019b)
88. S. Duncan, R. Carneiro, J. Braley, M. Hersh, F. Ramsey, R. Murch, Beyond ransomware: Securing the digital food chain (2021)
89. X.L. Palmer, E. Powell, L. Potter, Biocyberwarfare and crime: A juncture of rethought, in *European Conference on Cyber Warfare and Security* (Academic Conferences International Limited, 2021), pp. 517–XIV
90. R.J. Hester, Bioveillance: A techno-security infrastructure to preempt the dangers of informationalised biology. Sci. Culture **29**(1), 153–176 (2020)
91. K.M. Berger, P.A. Schneck, National and transnational security implications of asymmetric access to and use of biological data. Front. Bioeng. Biotechnol. **7**, 21 (2019)
92. J. Peccoud, J.E. Gallegos, R. Murch, W.G. Buchholz, S. Raman, Cyberbiosecurity: from naive trust to risk awareness. Trends Biotechnol. **36**(1), 4–7 (2018)
93. G. Turner, The growing need for cyberbiosecurity, in *InSITE 2019: Informing Science+ IT Education Conferences: Jerusalem* (2019), pp. 207–215
94. Q. Lin, S. Adepu, S. Verwer, A. Mathur, Tabor: A graphical model-based approach for anomaly detection in industrial control systems, in *Proceedings of the 2018 on Asia Conference on Computer and Communications Security* (2018), pp. 525–536
95. J.L. Mantle, J. Rammohan, E.F. Romantseva, J.T. Welch, L.R. Kauffman, J. McCarthy, J. Schiel, J.C. Baker, E.A. Strychalski, K.C. Rogers, et al., Cyberbiosecurity for biopharmaceutical products. Front. Bioeng. Biotechnol. **7**, 116 (2019)
96. C.O. Adetunji, O.T. Olugbemi, O.A. Anani, D.I. Hefft, N. Wilson, A.S. Olayinka, K.E. Ukhurebor, Cyberespionage: Socioeconomic implications on sustainable food security, in *AI, Edge and IoT-based Smart Agriculture* (Elsevier, 2022), pp. 477–486
97. D. Gillum, L.A.O. Carrera, I.A. Mendoza, P. Bates, D. Bowens, Z. Jetson, J. Maldonado, C. Mancini, M. Miraldi, R. Moritz, et al., The 2017 arizona biosecurity workshop: an open dialogue about biosecurity. Applied Biosafety **23**(4), 233–241 (2018)
98. L. Potter, X.L. Palmer, Human factors in biocybersecurity wargames, in *Future of Information and Communication Conference* (Springer, 2021), pp. 666–673
99. S. Adepu, A. Mathur, Introducing cyber security at the design stage of public infrastructures: A procedure and case study, in *Complex Systems Design & Management Asia* (Springer, 2016a), pp. 75–94
100. A. Ilyas, L. Engstrom, A. Athalye, J. Lin, Black-box adversarial attacks with limited queries and information, in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018* (2018). https://arxiv.org/abs/1804.08598
101. A. Hassanzadeh, A. Rasekh, S. Galelli, M. Aghashahi, R. Taormina, A. Ostfeld, M.K. Banks, A review of cybersecurity incidents in the water sector. J. Environ. Eng. **146**(5), 03120003 (2020)
102. F. Pasqualetti, F. Dörfler, F. Bullo, Attack detection and identification in cyber-physical systems. IEEE Trans. Automatic Control **58**(11), 2715–2729 (2013)
103. S. Hochreiter, J. Schmidhuber, Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)
104. J. Goh, S. Adepu, M. Tan, Z.S. Lee, Anomaly detection in cyber physical systems using recurrent neural networks, in *2017 IEEE 18th International Symposium on High Assurance Systems Engineering (HASE)* (IEEE, 2017), pp. 140–145
105. A.P. Mathur, N.O. Tippenhauer, Swat: A water treatment testbed for research and training on ics security, in *2016 International Workshop on Cyber-Physical Systems for Smart Water Networks (CySWater)* (IEEE, 2016), pp. 31–36
106. J. Inoue, Y. Yamagata, Y. Chen, C.M. Poskitt, J. Sun, Anomaly detection for a water treatment system using unsupervised machine learning, in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)* (IEEE, 2017), pp. 1058–1065
107. R. Taormina, S. Galelli, N.O. Tippenhauer, E. Salomons, A. Ostfeld, D.G. Eliades, M. Aghashahi, R. Sundararajan, M. Pourahmadi, M.K. Banks, et al., Battle of the attack detection algorithms: Disclosing cyber attacks on water distribution networks. J. Water Res. Plann. Manag. **144**(8), 04018048 (2018)

108. M. Aghashahi, R. Sundararajan, M. Pourahmadi, M.K. Banks, Water distribution systems analysis symposium–battle of the attack detection algorithms (batadal), in *World Environmental and Water Resources Congress 2017* (2017), pp. 101–108
109. B.M. Brentan, E. Campbell, G. Lima, D. Manzi, D. Ayala-Cabrera, M. Herrera, I. Montalvo, J. Izquierdo, E. Luvizotto Jr, On-line cyber attack detection in water networks through state forecasting and control by pattern recognition. in *World Environmental and Water Resources Congress 2017* (2017), pp. 583–592
110. S.E. Chandy, A. Rasekh, Z.A. Barker, B. Campbell, M.E. Shafiee, Detection of cyber-attacks to water systems through machine-learning-based anomaly detection in scada data, in *World Environmental and Water Resources Congress 2017* (2017), pp. 611–616
111. M. Giacomoni, N. Gatsis, A. Taha, Identification of cyber attacks on water distribution systems by unveiling low-dimensionality in the sensory data, in *World Environmental and Water Resources Congress 2017* (2017), pp. 660–675
112. M. Mardani, G. Mateos, G.B. Giannakis, Recovery of low-rank plus compressed sparse matrices with application to unveiling traffic anomalies. IEEE Trans. Inf. Theory **59**(8), 5186–5205 (2013)
113. A.A. Abokifa, K. Haddad, C.S. Lo, P. Biswas, Detection of cyber physical attacks on water distribution systems via principal component analysis and artificial neural networks, in *World Environmental and Water Resources Congress 2017* (2017), pp. 676–691
114. M.F.K. Pasha, B. Kc, S.L. Somasundaram, An approach to detect the cyber-physical attack on water distribution system, in *World Environmental and Water Resources Congress 2017* (2017), pp. 703–711
115. M. Housh, Z. Ohar, Integrating physically based simulators with event detection systems: Multi-site detection approach. Water Research **110**, 180–191 (2017)
116. M. Kravchik, A. Shabtai, Detecting cyber attacks in industrial control systems using convolutional neural networks, in *Proceedings of the 2018 Workshop on Cyber-Physical Systems Security and PrivaCy* (2018), pp. 72–83
117. M.A. Umer, A. Mathur, K.N. Junejo, S. Adepu, Generating invariants using design and data-centric approaches for distributed attack detection. Int. J. Crit. Infrastruct. Prot. **28**, 100341 (2020)
118. K.N. Junejo, J. Goh, Behaviour-based attack detection and classification in cyber physical systems using machine learning, in *Proceedings of the 2nd ACM International Workshop on Cyber-Physical System Security* (2016), pp. 34–43
119. S. Adepu, A. Mathur, Distributed detection of single-stage multipoint cyber attacks in a water treatment plant, in *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security* (2016), pp. 449–460
120. S. Adepu, A. Mathur, An investigation into the response of a water treatment system to cyber attacks, in *2016 IEEE 17th International Symposium on High Assurance Systems Engineering (HASE)* (IEEE, 2016), pp. 141–148
121. S. Adepu, A. Mathur, Distributed attack detection in a water treatment plant: Method and case study. IEEE Trans. Dependable Secure Comput. **18**(1), 86–99 (2018)
122. A. Al-Abassi, H. Karimipour, A. Dehghantanha, R.M. Parizi, An ensemble deep learning-based cyber-attack detection in industrial control system. IEEE Access **8**, 83965–83973 (2020)
123. M. Sermesant, H. Delingette, H. Cochet, P. Jaïs, N. Ayache, Applications of artificial intelligence in cardiovascular imaging. Nat. Rev. Cardiol. **18**(8), 600–609 (2021)
124. P. Sinčak, J. Ondo, D. Kaposztasova, M. Virčikova, Z. Vranayova, J. Sabol, Artificial intelligence in public health prevention of legionelosis in drinking water systems. Int. J. Environ. Res. Public Health **11**(8), 8597–8611 (2014)
125. J.M. Wing, Trustworthy AI. Commun. ACM **64**(10), 64–71 (2021)
126. S. Thiebes, S. Lins, A. Sunyaev, Trustworthy artificial intelligence. Electronic Markets **31**(2), 447–464 (2021)
127. V. Morckel, K. Terzano, Legacy city residents' lack of trust in their governments: An examination of flint, michigan residents' trust at the height of the water crisis. J. Urban Aff. **41**(5), 585–601 (2019)

128. O. Inderwildi, C. Zhang, X. Wang, M. Kraft, The impact of intelligent cyber-physical systems on the decarbonization of energy. Energy Environ. Sci. **13**(3), 744–771 (2020)
129. C.S. Wickramasinghe, D.L. Marino, J. Grandio, M. Manic, Trustworthy AI development guidelines for human system interaction, in *2020 13th International Conference on Human System Interaction (HSI)* (IEEE, 2020), pp. 130–136
130. R. Kaasschieter. The "why" in building trust in AI (2020). https://www.capgemini.com/2020/09/the-why-in-building-trust-in-ai/#:~:text=Accountability2C%20transparency%2C%20fairness%2C%20etc,they%20will%20not%20buy%20it
131. N.A. Smuha, The eu approach to ethics guidelines for trustworthy artificial intelligence. Comput. Law Rev. Int. **20**(4), 97–106 (2019)
132. H. Liu, Y. Wang, W. Fan, X. Liu, Y. Li, S. Jain, Y. Liu, A.K. Jain, J. Tang, Trustworthy AI: A computational perspective. Preprint (2021). arXiv:2107.06641
133. E. Toreini, M. Aitken, K.P. Coopamootoo, K. Elliott, V.G. Zelaya, P. Missier, M. Ng, A. van Moorsel, Technologies for trustworthy machine learning: A survey in a socio-technical context. Preprint (2020). arXiv:2007.08911
134. B.W. Israelsen, N.R. Ahmed, "dave... i can assure you... that it's going to be all right..." a definition, case for, and survey of algorithmic assurances in human-autonomy trust relationships. ACM Comput. Surv. (CSUR) **51**(6), 1–37 (2019)
135. G. Bernieri, M. Conti, F. Turrin, Evaluation of machine learning algorithms for anomaly detection in industrial networks, in *2019 IEEE International Symposium on Measurements & Networking (M&N)* (IEEE, 2019), pp. 1–6
136. S.D. Anton, S. Kanoor, D. Fraunholz, H.D. Schotten, Evaluation of machine learning-based anomaly detection algorithms on an industrial modbus/tcp data set, in *Proceedings of the 13th International Conference on Availability, Reliability and Security* (2018), pp. 1–9
137. H. Wiemer, A. Dementyev, S. Ihlenfeldt, A holistic quality assurance approach for machine learning applications in cyber-physical production systems. Applied Sciences **11**(20), 9590 (2021)
138. B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, B. Zhou, Trustworthy AI: From principles to practices. Preprint (2021b). arXiv:2110.01167
139. J. Mökander, L. Floridi, Ethics-based auditing to develop trustworthy AI. Minds Mach. **31**(2), 323–327 (2021)
140. E. Daglarli, Explainable artificial intelligence (xai) approaches and deep meta-learning models for cyber-physical systems, in *Artificial Intelligence Paradigms for Smart Cyber-Physical Systems* (IGI Global, 2021), pp. 42–67
141. D. Kaur, S. Uslu, A. Durresi, Requirements for trustworthy artificial intelligence—a review, in *International Conference on Network-Based Information Systems* (Springer, 2020), pp. 105–115
142. C. Louisell, K. Heaslip, Securing the digitally managed water supply, in *World Environmental and Water Resources Congress 2020: Emerging and Innovative Technologies and International Perspectives* (American Society of Civil Engineers Reston, VA, 2020), pp. 1–11
143. J. Bergal, Florida hack exposes danger to water systems (2021). https://www.pewtrusts.org/en/research-and-analysis/blogs/stateline/2021/03/10/florida-hack-exposes-danger-to-water-systems
144. B. Kerstein, Israel thwarts major coordinated cyber-attack on its water infrastructure command and control systems (2020). https://www.algemeiner.com/2020/04/26/israel-thwarts-major-coordinated-cyber-attack-on-its-water-infrastructure-command-and-control-systems/
145. M. Taddeo, T. McCutcheon, L. Floridi, Trusting artificial intelligence in cybersecurity is a double-edged sword. Nat. Mach. Intell. **1**(12), 557–560 (2019)
146. N. Nicolaou, D.G. Eliades, C. Panayiotou, M.M. Polycarpou, Reducing vulnerability to cyber-physical attacks in water distribution networks, in *2018 international workshop on cyber-physical systems for smart water networks (CySWater)* (IEEE, 2018), pp. 16–19
147. A. Khaled, S. Ouchani, Z. Tari, K. Drira, Assessing the severity of smart attacks in industrial cyber-physical systems. ACM Trans. Cyber Phys. Syst. **5**(1), 1–28 (2020)

148. F. Pasqualetti, F. Dörfler, F. Bullo, Cyber-physical security via geometric control: Distributed monitoring and malicious attacks, in *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)* (IEEE, 2012), pp. 3418–3425
149. Y. Wu, H.N. Dai, H. Tang, Graph neural networks for anomaly detection in industrial internet of things. IEEE Internet Things J. (2021)
150. B. Siegel, Industrial anomaly detection: A comparison of unsupervised neural network architectures. IEEE Sens. Lett. **4**(8), 1–4 (2020)
151. L. Rosa, T. Cruz, M.B. de Freitas, P. Quitério, J. Henriques, F. Caldeira, E. Monteiro, P. Simões, Intrusion and anomaly detection for the next-generation of industrial automation and control systems. Future Gener. Comput. Syst. **119**, 50–67 (2021)
152. L.A. Maglaras, J. Jiang, Intrusion detection in scada systems using machine learning techniques, in *2014 Science and Information Conference* (IEEE, 2014), pp. 626–631
153. C.M. Ahmed, G.R. MR, A.P. Mathur, Challenges in machine learning based approaches for real-time anomaly detection in industrial control systems, in *Proceedings of the 6th ACM on Cyber-Physical System Security Workshop* (2020), pp. 23–29
154. J. Zhang, L. Pan, Q.L. Han, C. Chen, S. Wen, Y. Xiang, Deep learning based attack detection for cyber-physical system cybersecurity: A survey. IEEE/CAA J. Automat. Sin. **9**(3), 377–391 (2021)
155. Y. Luo, Y. Xiao, L. Cheng, G. Peng, D. Yao, Deep learning-based anomaly detection in cyber-physical systems: Progress and opportunities. ACM Comput. Surv. (CSUR) **54**(5), 1–36 (2021)
156. N. Kadosh, A. Frid, M. Housh, Detecting cyber-physical attacks in water distribution systems: One-class classifier approach. J. Water Resour. Plann. Manag. **146**(8), 04020060 (2020)
157. D.C.L. Sung, G.R. MR, A.P. Mathur, Design-knowledge in learning plant dynamics for detecting process anomalies in water treatment plants. Comput. Secur. **113**, 102532 (2022)
158. D. Garcia, V. Puig, J. Quevedo, Prognosis of water quality sensors using advanced data analytics: Application to the barcelona drinking water network. Sensors **20**(5), 1342 (2020)
159. R. Taormina, S. Galelli, Real-time detection of cyber-physical attacks on water distribution systems using deep learning, in *World Environmental and Water Resources Congress 2017* (2017), pp. 469–479
160. A.A. Abokifa, K. Haddad, C. Lo, P. Biswas, Real-time identification of cyber-physical attacks on water distribution systems via machine learning–based anomaly detection techniques. J. Water Resour. Plann. Manag. **145**(1), 04018089 (2019)
161. N. Neha, S. Priyanga, S. Seshan, R. Senthilnathan, V. Shankar Sriram, Sco-rnn: A behavioral-based intrusion detection approach for cyber physical attacks in scada systems, in *Inventive Communication and Computational Technologies* (Springer, 2020), pp. 911–919
162. J. Kim, J.H. Yun, H.C. Kim, Anomaly detection for industrial control systems using sequence-to-sequence neural networks, in *Computer Security* (Springer, 2019), pp. 3–18
163. S. Sapkota, A. Mehdy, S. Reese, H. Mehrpouyan, Falcon: Framework for anomaly detection in industrial control systems. Electronics **9**(8), 1192 (2020)
164. C.H. Yoong, J. Heng, Framework for continuous system security protection in swat, in *Proceedings of the 2019 3rd International Symposium on Computer Science and Intelligent Control* (2019), pp. 1–6
165. L.H.A. Reis, A. Murillo Piedrahita, S. Rueda, N.C. Fernandes, D.S. Medeiros, M.D. de Amorim, D.M. Mattos, Unsupervised and incremental learning orchestration for cyber-physical security. Trans. Emerg. Telecommun. Technol. **31**(7), e4011 (2020)
166. M. Gauthama Raman, N. Somu, A.P. Mathur, Anomaly detection in critical infrastructure using probabilistic neural network, in *International Conference on Applications and Techniques in Information Security* (Springer, 2019), pp. 129–141
167. S. Kim, W. Jo, T. Shon, Apad: autoencoder-based payload anomaly detection for industrial ioe. Appl. Soft Comput. **88**, 106017 (2020)
168. S.K. Alabugin, A.N. Sokolov, Applying of generative adversarial networks for anomaly detection in industrial control systems, in *2020 Global Smart Industry Conference (GloSIC)* (IEEE, 2020), pp. 199–203

169. D.D. Tiwari, S. Naskar, A.S. Sai, V.R. Palleti, Attack detection using unsupervised learning algorithms in cyber-physical systems, in *Computer Aided Chemical Engineering*, vol. 50 (Elsevier, 2021), pp. 1259–1264
170. W. Zhou, X.-m. Kong, K.-l. Li, X.-m. Li, L.-l. Ren, Y. Yan, Y. Sha, X.-y. Cao, X.-j. Liu, Attack sample generation algorithm based on data association group by gan in industrial control dataset. Computer Communications **173**, 206–213 (2021)
171. M.G. Raman, W. Dong, A. Mathur, Deep autoencoders as anomaly detectors: Method and case study in a distributed water treatment plant. Comput. Secur. **99**, 102055 (2020)
172. R. Taormina, S. Galelli, Deep-learning approach to the detection and localization of cyber-physical attacks on water distribution systems. J. Water Resour. Plann. Manag. **144**(10), 04018065 (2018)
173. H. Wijaya, M. Aniche, A. Mathur, Domain-based fuzzing for supervised learning of anomaly detection in cyber-physical systems, in *Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops* (2020), pp. 237–244
174. P. Schneider, K. Böttinger, High-performance unsupervised anomaly detection for cyber-physical system networks, in *Proceedings of the 2018 Workshop on Cyber-Physical Systems Security and Privacy* (2018), pp. 1–12
175. M. Elnour, N. Meskin, K.M. Khan, Hybrid attack detection framework for industrial control systems using 1d-convolutional neural network and isolation forest, in *2020 IEEE Conference on Control Technology and Applications (CCTA)* (IEEE, 2020), pp. 877–884
176. R. Alguliyev, Y. Imamverdiyev, L. Sukhostat, Hybrid deepgcl model for cyber-attacks detection on cyber-physical systems. Neural Comput. Appl. **33**(16), 10211–10226 (2021)
177. Z. Chen, D. Chen, X. Zhang, Z. Yuan, X. Cheng, Learning graph structures with transformer for multivariate time series anomaly detection in iot. IEEE Internet Things J. (2021)
178. Y. Chen, C.M. Poskitt, J. Sun, S. Adepu, F. Zhang, Learning-guided network fuzzing for testing cyber-physical system defences, in *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)* (IEEE, 2019), pp. 962–973
179. A. Meleshko, V. Desnitsky, I. Kotenko, Machine learning based approach to detection of anomalous data from sensors in cyber-physical water supply systems, in *IOP Conference Series: Materials Science and Engineering*, vol. 709 (IOP Publishing, 2020), p. 033034
180. P. Perrone, F. Flammini, R. Setola, Machine learning for threat recognition in critical cyber-physical systems, in *2021 IEEE International Conference on Cyber Security and Resilience (CSR)* (IEEE, 2021), pp. 298–303
181. S. Athalye, C. Mujeeb Ahmed, J. Zhou, Model-based cps attack detection techniques: Strengths and limitations, in *Security in Cyber-Physical Systems* (Springer, 2021), pp. 155–187
182. A. Robles-Durazno, N. Moradpoor, J. McWhinnie, G. Russell, Z. Tan, Newly engineered energy-based features for supervised anomaly detection in a physical model of a water supply system. Ad Hoc Networks **120**, 102590 (2021)
183. J. Sun, Z. Yang, Objsim: efficient testing of cyber-physical systems, in *Proceedings of the 4th ACM SIGSOFT International Workshop on Testing, Analysis, and Verification of Cyber-Physical Systems and Internet of Things* (2020), pp. 1–2
184. S. Alqurashi, H. Shirazi, I. Ray, On the performance of isolation forest and multi layer perceptron for anomaly detection in industrial control systems networks, in *2021 8th International Conference on Internet of Things: Systems, Management and Security (IOTSMS)* (IEEE, 2021), pp. 1–6
185. M. Balaji, S. Shrivastava, S. Adepu, A. Mathur, Super detector: An ensemble approach for anomaly detection in industrial control systems, in *International Conference on Critical Information Infrastructures Security* (Springer, 2021), pp. 24–43
186. A.N. Jahromi, H. Karimipour, A. Dehghantanha, K.K.R. Choo, Toward detection and attribution of cyber-attacks in iot-enabled cyber–physical systems. IEEE Internet Things J. **8**(17), 13712–13722 (2021)
187. M. Baptiste, F. Julien, S. Franck, Systematic and efficient anomaly detection framework using machine learning on public ics datasets, in *2021 IEEE International Conference on Cyber Security and Resilience (CSR)* (IEEE, 2021), pp. 292–297

188. T. Chalongvorachai, K. Woraratpanya, A data generation framework for extremely rare case signals. Heliyon **7**(8), e07687 (2021)
189. G.R. MR, N. Somu, A.P. Mathur, A multilayer perceptron model for anomaly detection in water treatment plants. Int. J. Crit. Infrastruct. Prot. **31**, 100393 (2020)
190. P.F. de Araujo-Filho, G. Kaddoum, D.R. Campelo, A.G. Santos, D. Macêdo, C. Zanchettin, Intrusion detection for cyber–physical systems using generative adversarial networks in fog environment. IEEE Internet Things J. **8**(8), 6247–6256 (2020)
191. F. Turrin, A. Erba, N.O. Tippenhauer, M. Conti, A statistical analysis framework for ics process datasets, in *Proceedings of the 2020 Joint Workshop on CPS&IoT Security and Privacy* (2020), pp. 25–30
192. G. Sebestyen, A. Hangan, Z. Czako, Anomaly detection in water supply infrastructure systems, in *2021 23rd International Conference on Control Systems and Computer Science (CSCS)* (IEEE, 2021), pp. 349–355
193. L. Gjorgiev, S. Gievska, Time series anomaly detection with variational autoencoder using mahalanobis distance, in *International Conference on ICT Innovations* (Springer, 2020), pp. 42–55
194. S. Chockalingam, W. Pieters, A. Teixeira, P. van Gelder, Bayesian network model to distinguish between intentional attacks and accidental technical failures: a case study of floodgates. Cybersecurity **4**(1), 1–19 (2021)
195. R. Qadeer, C. Murguia, C.M. Ahmed, J. Ruths, Multistage downstream attack detection in a cyber physical system, in *Computer Security* (Springer, 2017), pp. 177–185
196. C.M. Ahmed, S. Adepu, A. Mathur, Limitations of state estimation based cyber attack detection schemes in industrial control systems, in *2016 Smart City Security and Privacy Workshop (SCSP-W)* (IEEE, 2016), pp. 1–5
197. C.M. Ahmed, M. Ochoa, J. Zhou, A.P. Mathur, R. Qadeer, C. Murguia, J. Ruths, Noiseprint: Attack detection using sensor and process noise fingerprint in cyber physical systems, in *Proceedings of the 2018 on Asia Conference on Computer and Communications Security* (2018), pp. 483–497
198. T.K. Das, S. Adepu, J. Zhou, Anomaly detection in industrial control systems using logical analysis of data. Comput. Secur. **96**, 101935 (2020)
199. S. Adepu, J. Prakash, A. Mathur, Waterjam: An experimental case study of jamming attacks on a water treatment system, in *2017 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C)* (IEEE, 2017), pp. 341–347
200. S. Liyakkathali, F. Furtado, G. Sugumar, A. Mathur, A mechanism to assess the effectiveness anomaly detectors in industrial control systems. J. Integr. Des. Process Sci. (Preprint), 1–26 (2022)
201. G. Sugumar, A. Mathur, Testing the effectiveness of attack detection mechanisms in industrial control systems, in *2017 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C)* (IEEE, 2017), pp. 138–145
202. A. Mathur, Secwater: A multi-layer security framework for water treatment plants, in *Proceedings of the 3rd International Workshop on Cyber-Physical Systems for Smart Water Networks* (2017), pp. 29–32
203. D. Dovžan, V. Logar, I. Škrjanc, Implementation of an evolving fuzzy model (efumo) in a monitoring system for a waste-water treatment process. IEEE Trans. Fuzzy Syst. **23**(5), 1761–1776 (2014)
204. S. Adepu, S. Shrivastava, A. Mathur, Argus: An orthogonal defense framework to protect public infrastructure against cyber-physical attacks. IEEE Internet Comput. **20**(5), 38–45 (2016)
205. S. Adepu, A. Mathur, Assessing the effectiveness of attack detection at a hackfest on industrial control systems. IEEE Trans. Sustain. Comput. **6**(2), 231–244 (2018b)
206. D. Urbina, J. Giraldo, N.O. Tippenhauer, A. Cardenas, Attacking fieldbus communications in ics: Applications to the swat testbed, in *Proceedings of the Singapore Cyber-Security Conference (SG-CRC) 2016* (IOS Press, 2016), pp. 75–89

207. K. Pal, S. Adepu, J. Goh, Effectiveness of association rules mining for invariants generation in cyber-physical systems, in *2017 IEEE 18th International Symposium on High Assurance Systems Engineering (HASE)* (IEEE, 2017), pp. 124–127
208. M.A. Umer, A. Mathur, K.N. Junejo, S. Adepu, Integrating design and data centric approaches to generate invariants for distributed attack detection, in *Proceedings of the 2017 Workshop on Cyber-Physical Systems Security and Privacy* (2017), pp. 131–136
209. E. Kang, S. Adepu, D. Jackson, A.P. Mathur, Model-based security analysis of a water treatment system, in *2016 IEEE/ACM 2nd International Workshop on Software Engineering for Smart Cyber-Physical Systems (SEsCPS)* (IEEE, 2016), pp. 22–28
210. S. Shrivastava, G.R. MR, A. Mathur, Pcat: Plc command analysis tool for automatic incidence response in water treatment plants, in *2021 IEEE International Conference on Big Data (Big Data)* (IEEE, 2021), pp. 2151–2159
211. A. Robles-Durazno, N. Moradpoor, J. McWhinnie, G. Russell, I. Maneru-Marin, Plc memory attack detection and response in a clean water supply system. Int. J. Crit. Infrastruct. Prot. **26**, 100300 (2019)
212. A. Agrawal, C.M. Ahmed, E.C. Chang, Poster: Physics-based attack detection for an insider threat model in a cyber-physical system, in *Proceedings of the 2018 on Asia Conference on Computer and Communications Security* (2018), pp. 821–823
213. N. Chikhalia, Y. Dhawan, Security of industrial cyberspace: Fair clustering with linear time approximation, in *Handbook of Big Data Analytics and Forensics* (Springer, 2022), pp. 75–88
214. J. Wang, J. Sun, Y. Jia, S. Qin, Z. Xu, Towards 'verifying' a water treatment system, in *International Symposium on Formal Methods* (Springer, 2018), pp. 73–92
215. G.R. MR, A.P. Mathur, Aicrit: A unified framework for real-time anomaly detection in water treatment plants. J. Inf. Secur. Appl. **64**, 103046 (2022)
216. S. Uslu, D. Kaur, S.J. Rivera, A. Durresi, M. Babbar-Sebens, J.H. Tilt, A trustworthy human–machine framework for collective decision making in food–energy–water management: The role of trust sensitivity. Knowl. Based Syst. **213**, 106683 (2021)
217. K. Krithivasan, S. Pravinraj, V.S. Shankar Sriram, et al., Detection of cyberattacks in industrial control systems using enhanced principal component analysis and hypergraph-based convolution neural network (epca-hg-cnn). IEEE Trans. Ind. Appl. **56**(4), 4394–4404 (2020)
218. Q. Xu, S. Ali, T. Yue, Digital twin-based anomaly detection in cyber-physical systems, in *2021 14th IEEE Conference on Software Testing, Verification and Validation (ICST)* (IEEE, 2021), pp. 205–216
219. Z. Li, J. Li, Y. Wang, K. Wang, A deep learning approach for anomaly detection based on sae and lstm in mechanical equipment. Int. J. Adv. Manuf. Technol. **103**(1), 499–510 (2019)
220. M.F. Abdelaty, R.D. Corin, D. Siracusa, Daics: A deep learning solution for anomaly detection in industrial control systems. IEEE Trans. Emerg. Top. Comput. (2021)
221. M. Elnour, N. Meskin, K. Khan, R. Jain, A dual-isolation-forests-based attack detection framework for industrial control systems. IEEE Access **8**, 36639–36651 (2020)
222. M. Kravchik, A. Shabtai, Efficient cyber attack detection in industrial control systems using lightweight neural networks and pca. IEEE Trans. Dependable Secure Comput. (2021)
223. Á.L. Perales Gómez, L. Fernández Maimó, A. Huertas Celdrán, F.J. García Clemente, Madics: A methodology for anomaly detection in industrial control systems. Symmetry **12**(10), 1583 (2020)
224. G. Zizzo, C. Hankin, S. Maffeis, K. Jones, Intrusion detection for industrial control systems: Evaluation analysis and adversarial attacks. Preprint (2019). arXiv:1911.04278
225. D. Shalyga, P. Filonov, A. Lavrentyev, Anomaly detection for water treatment system based on neural network with automatic architecture optimization. Preprint (2018). arXiv:1807.07282
226. D. Li, D. Chen, B. Jin, L. Shi, J. Goh, S.K. Ng, Mad-gan: Multivariate anomaly detection for time series data with generative adversarial networks, in *International Conference on Artificial Neural Networks* (Springer, 2019), pp. 703–716

227. K. Faber, M. Pietron, D. Zurek, Ensemble neuroevolution-based approach for multivariate time series anomaly detection. Entropy **23**(11), 1466 (2021)
228. S. Chakraborty, A. Onuchowska, S. Samtani, W. Jank, B. Wolfram, Machine learning for automated industrial iot attack detection: an efficiency-complexity trade-off. ACM Trans. Manag. Inf. Syst. (TMIS) **12**(4), 1–28 (2021)
229. D. Li, D. Chen, J. Goh, S.k. Ng, Anomaly detection with generative adversarial networks for multivariate time series. Preprint (2018). arXiv:1809.04758
230. B. Brentan, P. Rezende, D. Barros, G. Meirelles, E. Luvizotto, J. Izquierdo, Cyber-attack detection in water distribution systems based on blind sources separation technique. Water **13**(6), 795 (2021)
231. D.T. Ramotsoela, G.P. Hancke, A.M. Abu-Mahfouz, Attack detection in water distribution systems using machine learning. HCIS **9**(1), 1–22 (2019)
232. L. Tsiami, C. Makropoulos, Cyber-physical attack detection in water distribution systems with temporal graph convolutional neural networks. Water **13**(9), 1247 (2021)
233. H. Mahmoud, W. Wu, M.M. Gaber, A time-series self-supervised learning approach to detection of cyber-physical attacks in water distribution systems. Energies **15**(3), 914 (2022)
234. Y.H. Choi, A. Sadollah, J.H. Kim, Improvement of cyber-attack detection accuracy from urban water systems using extreme learning machine. Applied Sciences **10**(22), 8179 (2020)
235. B.A. Vinatzer, L.S. Heath, H.M. Almohri, M.J. Stulberg, C. Lowe, S. Li, Cyberbiosecurity challenges of pathogen genome databases. Front. Bioeng. Biotechnol. **7**, 106 (2019)
236. J. Diggans, E. Leproust, Next steps for access to safe, secure dna synthesis. Front. Bioeng. Biotechnol. **7**, 86 (2019)
237. R. Puzis, D. Farbiash, O. Brodt, Y. Elovici, D. Greenbaum, Increased cyber-biosecurity for DNA synthesis. Nature Biotechnology **38**(12), 1379–1381 (2020)
238. D. Farbiash, R. Puzis, Cyberbiosecurity: Dna injection attack in synthetic biology. Preprint (2020). arXiv:2011.14224
239. S. Mueller, On DNA signatures, their dual-use potential for gmo counterfeiting, and a cyber-based security solution. Front. Bioeng. Biotechnol. **7**, 189 (2019)
240. D. Guttieres, S. Stewart, J. Wolfrum, S.L. Springs, Cyberbiosecurity in advanced manufacturing models. Front. Bioeng. Biotechnol., 210 (2019)
241. Z. Li, H. Zhao, J. Shi, Y. Huang, J. Xiong, An intelligent fuzzing data generation method based on deep adversarial learning. IEEE Access **7**, 49327–49340 (2019)
242. P. Rana, L.R. Varshney, Trustworthy predictive algorithms for complex forest system decision-making. Front. Forests Global Change, 153 (2021)
243. A.M. George, The national security implications of cyberbiosecurity. Front. Bioeng. Biotechnol. **7**, 51 (2019)
244. X.L. Palmer, S. Karahan, Defending forward: an exploration through the lens of biocybersecurity, in *ICCWS 2020 15th International Conference on Cyber Warfare and Security* (Academic Conferences and Publishing Limited, 2020), p. 373
245. X.L. Palmer, L. Potter, S. Karahan, On the emerging area of biocybersecurity and relevant considerations, in *Future of Information and Communication Conference* (Springer, 2020), pp. 873–881
246. A.F. Markus, J.A. Kors, P.R. Rijnbeek, The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. J. Biomed. Inf. **113**, 103655 (2021)
247. A. Holzinger, C. Biemann, C.S. Pattichis, D.B. Kell, What do we need to build explainable AI systems for the medical domain? Preprint (2017). arXiv:1712.09923
248. M. Quiñones-Grueiro, A. Prieto-Moreno, C. Verde, O. Llanes-Santiago, Decision support system for cyber attack diagnosis in smart water networks. IFAC-PapersOnLine **51**(34), 329–334 (2019)
249. S. Adepu, A. Mathur, Using process invariants to detect cyber attacks on a water treatment system, in *IFIP International Conference on ICT Systems Security and Privacy Protection* (Springer, 2016), pp. 91–104

250. M. Macas, C. Wu, An unsupervised framework for anomaly detection in a water treatment system, in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)* (IEEE, 2019), pp. 1298–1305
251. A. Deng, B. Hooi, Graph neural network-based anomaly detection in multivariate time series, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35 (2021), pp. 4027–4035
252. C. Gehrmann, M. Gunnarsson, A digital twin based industrial automation and control system security architecture. IEEE Trans. Ind. Inf. **16**(1), 669 (2019)
253. Y. Jia, J. Wang, C.M. Poskitt, S. Chattopadhyay, J. Sun, Y. Chen, Adversarial attacks and mitigation for anomaly detectors of cyber-physical systems. Int. J. Crit. Infrastruct. Prot. **34**, 100452 (2021)
254. J.H. Moon, J.H. Yu, K.A. Sohn, An ensemble approach to anomaly detection using high-and low-variance principal components. Comput. Electr. Eng. **99**, 107773 (2022)
255. R.S. Murch, W.K. So, W.G. Buchholz, S. Raman, J. Peccoud, Cyberbiosecurity: an emerging new discipline to help safeguard the bioeconomy. Front. Bioeng. Biotechnol., 39 (2018)
256. D.S. Schabacker, L.A. Levy, N.J. Evans, J.M. Fowler, E.A. Dickey, Assessing cyberbiosecurity vulnerabilities and infrastructure resilience. Front. Bioeng. Biotechnol. **7**, 61 (2019)
257. K. Demestichas, N. Peppes, T. Alexakis, Survey on security threats in agricultural iot and smart farming. Sensors **20**(22), 6458 (2020)
258. S. Gurrapu, F.A. Batarseh, P. Wang, M.N.K. Sikder, N. Gorentala, M. Gopinath, Deepag: Deep learning approach for measuring the effects of outlier events on agricultural production and policy. in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)* (IEEE, 2021), pp. 1–8
259. M. Gopinath, F.A. Batarseh, J. Beckman, Machine learning in gravity models: An application to agricultural trade. Tech. rep., National Bureau of Economic Research (2020)
260. A. Monken, F. Haberkorn, M. Gopinath, L. Freeman, F.A. Batarseh, Graph neural networks for modeling causality in international trade, in *The International FLAIRS Conference Proceedings*, vol. 34 (2021)