The Challenges and Opportunities of Human-Centred AI for Trustworthy Robots and Autonomous Systems

Hongmei He, *Senior Member, IEEE*, John Gray, *Member, IEEE*,
Angelo Cangelosi, *Senior Member, IEEE,* Qinggang Meng, *Senior Member*, *IEEE*,
T.M. McGinnity, *Senior Member*, *IEEE,* Jörn Mehnen, *Member, IEEE*

Abstract— **The trustworthiness of robots and autonomous systems (RAS) has taken a prominent position on the way towards full autonomy. This work is the first to systematically explore the key facets of human-centred AI for trustworthy RAS. We identified five key properties of a trustworthy RAS, i.e., RAS must be (i) safe in any uncertain and dynamic environment; (ii) secure, i.e., protect itself from cyber threats; (iii) healthy and fault-tolerant; (iv) trusted and easy to use to enable effective human-machine interaction (HMI); (v) compliant with the law and ethical expectations. While the applications of RAS have mainly focused on performance and productivity, not enough scientific attention has been paid to the risks posed by advanced AI in RAS. We analytically examine the challenges of implementing trustworthy RAS with respect to the five key properties and explore the role and roadmap of AI technologies in ensuring the trustworthiness of RAS in respect of safety, security, health, HMI, and ethics. A new acceptance model of RAS is provided as a framework for human-centric AI requirements and for implementing trustworthy RAS by design. This approach promotes human-level intelligence to augment human capabilities and focuses on contribution to humanity.**

***Index Terms*—Human-centred Artificial Intelligence, Trustworthiness of RAS, Cyber Security, Safety, System Health, Human-Robot Interaction, Performance of RAS, Acceptance Model, Trustiness, Worthiness.**

I. Introduction

Robots and Autonomous Systems (RAS), equipped with Artificial Intelligence (AI) technologies are attempting to mimic the adaptive and smart capabilities of human problem solving [1], and to enhance the abilities to perceive complex environments and make decisions quickly by stimulating cognitive and developmental abilities toward human intelligence. Machine learning (ML) is developed out of the quest for AI, and the core objective of a learner is to create a general target function from its experience to map the relationships between inputs and outputs. Therefore, ML provides a technical approach to solving those problems that are not programmable. The introduction to ML promoted the development of AI. RAS allow for creating smart and safe work environments where humans are relieved from the burden of arduous, repetitive or dangerous tasks. RAS can also help where super-human quick and precise actions are of the essence. For example, driverless cars equipped with intelligent safe road assist systems can greatly reduce the frequency of road accidents, medical robots with smart augmented reality technology enhance the performance of intricate surgeries, and intelligent autopilots perform delicate docking manoeuvres.

The vast benefits of RAS have made this technology popular in many application domains such as aerospace, transport, manufacturing, agriculture, social healthcare, and extreme environments. The spectrum of RAS applications spans e.g. robotics, smart factories, autonomous vehicles, unmanned aerial vehicles, autonomous trading systems, self-managing telecommunication networks, and smart infrastructure.

The Internet of Things (IoT) delivers new value by connecting people, processes and data. Sensing and data analysis technologies in IoT are giving robots a wider situational awareness that leads to better task execution. The concept of the Internet of Robotic Things (IoRT), raised by ABI research [2], introduces robots into the IoT application domains, creating harmonic collaboration between human, machines, and the physical world. IoRT technology extends the application scope of RAS and makes it an extremely powerful tool.

It is impossible to anticipate all potential challenging situations that a RAS may experience in practice. Advanced RAS must behave robustly and safely in any critical situation. Trust is built on predictability and understanding. Therefore, Trustworthy RAS (TRAS) by design must be the starting point to ensure progress towards trustworthy, fully autonomous systems.

H. He is with the School of Computer Science and Informatics, De Montfort University, Leicester, UK, LE1 9BH (e-mail: h.he@dmu.ac.uk).
J. Gray is with the Department of Electronics and Electrical Engineering, the University of Manchester, Manchester, UK. M13 9PL (e-mail: john.gray-2@manchester.ac.uk).
A. Cangelosi is the Department of Computer Science, the University of Manchester, Manchester, UK, M13 9PL (e-mail: angelo.cangelosi@manchester.ac.uk).
Q. Meng is with the Department of Computer Science, Loughborough University, Loughborough, UK, LE11 3TU (e-mail: q.meng@lboro.ac.uk).
T.M. McGinnity is with the Department of Computer Science, Nottingham Trent University, Nottingham, UK, NG1 4FQ (e-mail: martin.mcginnity@ntu.ac.uk) and also with the Intelligent Systems Research Centre, Ulster University (tm.mcginnity@ulster.ac.uk)
J. Mehnen is with Design, Manufacturing and Eng. Management, University of Strathclyde, Glasgow, UK, G1 1XQ (e-mail: jorn.mehnen@strath.ac.uk).

This is a peer-reviewed, accepted author manuscript of the following article: He, H., Gray, J., Cangelosi, A., Meng, Q., McGinnity, T. M., & Mehnen, J. (2021). The challenges and opportunities of human-centred AI for trustworthy robots and autonomous systems. IEEE Transactions on Cognitive and Developmental Systems. https://doi.org/10.1109/TCDS.2021.3132282

The trustworthiness of AI is one of the top AI challenges today. As time and technology progress, it becomes increasingly evident that these systems must be trustworthy to reach the plateau of productivity. With regards to autonomous systems, hard lessons have already been learned. For example, a self-driving car hit a lady in Arizona in 2018 in a fatal accident [3]; hijacking a car on a motorway proved successfully that security flaws exist in a modern remote accessible vehicle [4]; two airplane accidents that cost 346 human lives were caused by failures of the onboard intelligent aviation systems [5]. Such accidents damage the public trust in RAS technologies.

The examples above demonstrate that trustworthy RAS requires addressing ethical, societal, regulatory, and educational challenges as well [6]. One of the most common ethical examples is that autonomous systems may face the so-called "trolley problem" [7], which involves stylized ethical dilemmas of whether to sacrifice one person to save others. While some are regarding the incorporation of trustworthiness in RAS only as an additional cost burden, in September 2020 the European Parliament estimated in its latest European Value-Added Assessment that a joint EU approach to ethical aspects of AI can add an extra €294 billion in GDP and 4.6 million jobs in the EU by 2030 [8].

In this paper, we address the relevant technological and ethical properties for implementing trustworthy RAS. We explore the challenges and opportunities of AI in implementing trustworthy RAS. We also present a model for human acceptance of RAS and shed light on human-centred AI as we move towards fully autonomous systems. UKRI is investing £33 million in a Trustworthy Autonomous Systems (TAS) program [9]. The research presented here is fully aligned with UKRI's TAS program.

The paper is organized as follows. Section II identifies key factors affecting the trustworthiness of RAS; Section III addresses the challenges of implementing trustworthy RAS in terms of the key properties; Section IV explores the role of AI in implementing trustworthy RAS; Section V proposes a human acceptance mode for RAS and addresses the importance of human-centred AI, and finally Section VI concludes the work.

II. Key Factors That Affect TRAS

The growing concern over the need for trustworthy RAS has led to some initial international efforts to develop approaches to ensure and enhance trustworthiness. In the USA, the National Institute of Standards and Technology (NIST) has developed a framework encapsulating the key aspects of Cyber Physical Systems (CPS) [10], as shown in Fig. 1. The NIST CPS model can be applied to RAS because most types of RAS belong to CPS, even if a fully autonomous system requires neither human intervention nor supporting continuous Internet connectivity. RAS is performed by a closed loop of four phases of CPS plus HMI. However, cyber-attacks could penetrate from upper layer IT infrastructure to the physical layer, and directly influence the decisions and actions of the physical layer. Therefore, the physical state in the loop should be extended to consider environment, system and cyber states. In the following, we explore critical factors of trustworthy RAS that can contribute to increase public acceptance of RAS.

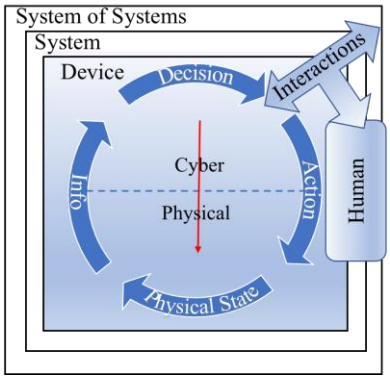

Fig. 1. The NIST CPS Framework Release 1.0 [10]

A robot or autonomous system is considered trustworthy only if it can demonstrate reliable and safe performance in its core functions. For most RAS these characteristics begin with the system's sensing and data acquisition modalities upon which the unit's perception, decision making, communication, and actions are built. Logic, reliability, correctness, and transparency of decision making, performance, and overall quality are essential requirements for autonomous systems on which trust is built.

Safety is an essential requirement for all types of RAS. Different applications of RAS may require different levels of safety. In aerospace, the safety of passengers and crew is crucial; in the automotive industry, the safety of vehicle users, other road users, and surrounding facilities is crucial. To ensure safety and monitor the environment, modern cars increasingly use advanced technologies such as smart cameras, GPS, radar, LiDAR and other types of sensors, as well as on-board computers. Conversely, while the safety of an autonomous household appliance, such as a robotic vacuum cleaner, is important, a failure would be much less likely to have a catastrophic effect. Safety in Human-Robot Interaction (HRI) is critical in modern AI-equipped robotic communities. Cognition-based robot control in HRI, action planning for safe navigation, hardware safety features, and social and psychological factors can positively contribute to trustworthy HRI [11][12]. Safety is directly related to the reliability of RAS, which is an important factor when a human selects a RAS.

Security has become the focus of user concern due to high-visibility data breaches in major companies. However, data breaches are only one aspect of the security of RAS. A great challenge is the security for cyber-physical systems where cyber-attacks pose the risk not only of information loss but also of direct takeover of control of an entity. A classic example is the cyberattack on the Jeep SUV in 2015 [5]. This could result in devastating consequences and damage the safety of RAS. A trustworthy RAS should protect personal data and the system from cyber-attacks while ensuring regulatory compliance (e.g., GDPR). Considerable research efforts have been put into the development of security-related robotic solutions equipped with AI methods [13][14].

Internal System Health. In addition to the security challenges, the reliability of RAS faces the challenges of any electronic/computational system, namely component failures and unanticipated behaviours. Various faults in RAS occur due to erroneous sensor measurements and incomplete sensor data,

or errors in actuator movement. Most difficult faults arise with respect to the underlying AI system due to misinterpretation of sensor inputs or actuator feedback, which can lead to erroneous perceptions and catastrophic decisions. Fault diagnosis is not a new problem that the electronics industry has been dealing with for decades, but its importance has increased as systems have become more complex and autonomous. It is critical that reliability, reliable degradation, and redundancy play a central role in the design of RAS to ensure early detection of various potential anomalies. It is crucial that the design of RAS incorporates fault-tolerant operations to reduce performance degradation or provides pre-emptive notification of impending errors to reduce the potential for dangerous situations [15].

Human-Machine Interaction (HMI) refers to the bidirectional interactions between the machine and the human user. HMI occurs through a user interface based on human sensory modalities - speech, touch, vision, and possibly smell. An important consideration is whether the machine is considered an equal partner in the decision loop or whether the human has the right to override the machine's decisions. In the NISP's CPS model (Fig. 1), human input can be fed into the decision loop. In principle, this is correct, but two major problems arise: (a) the complexity of the decision process, which affects the quality and pertinence of the human input, and (b) the problem of interrupting a RAS in the middle of a critical process. Regarding a trustworthy system, it is critical that the human is allowed to interact with or interrupt the system. Such an interruption must be done in a way that does not increase the danger or exacerbate an error and must be based on efficient human-machine communication that takes into account the state of the art in human-machine interface implementation. Such systems incorporate advanced visualisation that allows humans to quickly grasp and interpret the current and predicted states of the system, determine developmental trends, or estimate dynamics, and thereby maintain consistency of systems in their dynamic operating environment. As a minimum, the principles of "Design for Error" [16] should assume that faults will occur and provide a reasonable margin for human intervention, where an "error" is broadly understood as a set of circumstances that cause the machine to deviate from the decision trajectory expected by humans**.**

Ethics is of increasing public concern in the way of implementing trustworthy autonomous systems. To date, the moral standing and dignity of a human cannot be transferred to a machine. Although the vast majority of recent developments on RAS improved people's lives and had a positive impact on the society, their continued rapid development raises significant ethical issues. Kallioinen et al. [17] investigated the ethical issues of autonomous vehicles by conducting virtual experiments on some scenarios, such as child pedestrians versus adult pedestrians, pedestrians on the street versus pedestrians on the sidewalk, and car occupants versus pedestrians. The experimental results show that human drivers and self-driving cars were largely judged similarly. When faced with a non-deterministic problem, an intelligent system can be designed to make a decision based on the probabilities of multiple states in terms of conventional rules. It is crucial, of course, that the final decision follows the relevant regulation rather than an engineers' preference. The immediate problem is the speed of developments in AI and robotics, which by far outstrips the pace at which the necessary accompanying ethical and moral framework are developed. There is an urgent need to agree common international standards, frameworks and guidelines to inform RAS regulations and relevant legislations.

III. The Challenges in Implementing TRAS

Optimising the performance of RAS faces significant challenges due to the diversity, uncertainty, and complexity of tasks in different application domains. The main challenge is that the performance and quality of RAS for different tasks in the context of a specific application domain must satisfy the five properties of a trusted RAS.

A. *Challenges for RAS Safety*

Regarding the safety of RAS, two aspects need to be considered: the threats to the safety of RAS from the environment and threats to the safety of users or assets in the environment. They are mainly determined by the perception technologies and response behaviours of RAS. As shown in Fig. 1, the decision system receives the information from the sensing system or the operator and creates action plans according to an abstract representation of the system and its environment. For example, in a robot navigation system, the decision output is the optimal path for the robot to avoid all obstacles and reach the destination. Thus, for safe navigation, the robot must be able to effectively and efficiently detect obstacles on the way to the destination [18]. To enable RAS to operate safely under different circumstances, the following challenges must be overcome:

(1) *Dynamic environments with uncertainty*. RAS must be able to correctly sense its environment, make real-time decisions in exceptional or emergency situations, and react quickly without deviating from the goal set for RAS.

(2) *Real-time synchronisation*. RAS must be able to safely perform various tasks, work harmoniously with humans and other robots in a collaborative team, and avoid accidents that could harm other team members and themselves.

(3) *Dealing with unexpected errors*. RAS must be able to effectively and efficiently detect unknown errors, handle all exceptions, and provide timely warnings of and prompt responses to system errors.

(4) *The ability to act safely in a situation for which RAS has no prior experience or pre-programmed response*. RAS should have the capability of self-learning, thus handling some unexpected events to reduce risks that threaten the safety of users, themselves or others. However, it might not survive all unexpected events that occur for the first time. Therefore, timely human intervention in an emergent situation at any execution time is crucial.

B. *Challenges of RAS Security*

With the advent of IoRT to address challenges in real-world applications through the use of sensors, artificial intelligence, software and communication technologies, etc., cyber threats are shifting from IT infrastructure in the digital world to actuation systems in the physical world [19]. As a result, the attack surface for IoRT systems has increased significantly. Cyber-attacks or cybercrimes not only threaten known devices in the IT infrastructure at the upper layers of the IoT stack, but

also target conventional communication protocols and RAS at the lower layers of the IoT stack. Due to RAS connectivity, more access points are potentially vulnerable to cyber-attacks, and through these attack points, attackers can penetrate a system, e.g., by injecting data into the system or extracting data from the system, thus compromising the security and control.

Usually, researchers paid mostly attention to the cybersecurity of hardware and software of autonomous systems and communication between different devices [20]. However, there are few studies on cyber threats to the development platforms of RAS, which directly affect the security of the systems being developed.

The first challenge for the security of RAS is to secure the communication links. For example, public transportation (e.g., buses and trains) provide Wi-Fi services; GPS, radio and Bluetooth have been used in modern vehicles; 5G communication between vehicles and road infrastructure is a future trend for connected and autonomous vehicles. However, different communication channels provide the opportunity to penetrate a RAS using different attack techniques, such as Trojan attacks on cryptographic protocols or via the quantum channel [21], man-in-the-middle attacks between vehicle-to-vehicle communication, MAC spoofing, wireless hijacking, denial of services (DoS), malicious eavesdropping [22], and attacks that exploit the vulnerability of Key Negotiation of Bluetooth (KNOB) [23].

The second challenge is to ensure the integrity of the RAS software. Integrity is one of the most important security goals in the security triangle (confidentiality, integrity, and availability). System software integrity requires protecting RAS software from code modification, malfunction, loss of control, and loss of personally identifiable information to communication loss and network congestion. Any disruption to the integrity of software can have serious consequences. Loss of control and malfunctions are particularly critical, which can compromise the safety of RAS and even endanger the lives [24].

The third critical challenge is securing the hardware. An autonomous vehicle consists of various components, including mechanical and electronic components, especially many Embedded Computing Units (ECUs) that could serve as attack points. The resource constraints of embedded systems lead to tight limits on communication and computing capacity. These constraints make it difficult to develop advanced security solutions for embedded systems [25]. The hardware attack surface can be any component in RAS such as sensors, USB ports or input/output units, and embedded systems. If an attacker penetrates the sensor system of RAS, the data provided by the sensors could lead to a wrong decision, which in turn results in a wrong action of the system. Typical side-channel attacks on embedded systems that steal secret information without leaving a trace on the device include fault injection attacks, energy analysis attacks, timing analysis attacks, and electromagnetic analysis attacks [26]. For example, a Rowhammer fault injection attack can be performed remotely to gain complete access to a device DRAM (Dynamic Random-Access Memory) and cache side-channel attacks obtain secret information by monitoring the system's cache [27].

C. *Challenges to RAS Health*

The document ISO 10303-226 [28] defines a fault as an abnormal condition or defect at the component, device, or subsystem level that results in failure. Failures can be classified into three categories: deterioration (fatigue), sudden failure (noise), and initial failure [29]. Diagnosis is the ability to detect, isolate, and identify which component has a fault and analyse the potential impact of the failed component on the state of the system, while prediction is the ability to predict upcoming states of a system or a fault before it occurs [30]. RAS places higher demands on fault prediction than other systems because a fault in an autonomous system could have more severe consequences. It is challenging to obtain highly accurate prediction information because this is highly dependent on system usage, operator experience, and work environment. Uncertainty factors have a direct impact on the fatigue level and speed of the system components, however, it is difficult to quantify these factors.

Traditional hazard analysis methods include Fault Tree Analysis (FTA) [31], Hazard and Operability Analysis (HAZOP) [32], Failure Modes, Effects and Criticality Analysis (FMECA) [33] and Systems Theoretic Process Analysis (STPA) [34]. These methods can precisely identify the potential risks and model fault scenarios during the design phase. but it is still almost impossible to incorporate these methods in the operation phase of a system. Therefore, online self-diagnosis is needed.

The self-diagnostic system of RAS mainly includes three functions: internal state detection, fault diagnosis and tolerance, and fast responses to sudden faults [29]. The complexity, variety, and uncertainty of faults present many challenges. With complicated faults, it is particularly difficult to find the cause of the fault, which in turn brings the challenge of reacting quickly, correctly and in real time.

D. *Challenges of Human-RAS Interaction*

Regardless of the level of autonomy at which a system is implemented, HMI is essential in the system. Even if it is a fully autonomous system, human-machine interaction is still essential because humans must keep control of the system in case of any emergency. This is in line with the modern view that humans are complemented by AI and not replaced by AI and RAS. The studies of Veloso [35] have shown that it is difficult to interrupt a robot during its autonomous execution unless we provide a proper interruption mechanism or turn off the power. However, an unexpected power shutdown can damage the system. It can be difficult to abruptly interrupt a robot safely at any execution point. For this, we need to go through all scenarios where different temporal-spatial parameters and constraints for different tasks, such as task priorities, operations, interruption frequency, and timing, make it challenging to implement an effective and efficient HMI.

Many applications require humans and robots to work together. In safety-critical domains (e.g., defense, healthcare, and aerospace), the consequences of misoperation, mis-response, and failure can be extremely severe, resulting in major economic losses and even tragic human casualties. One of the advantages of using RAS is that RAS could help with performing dangerous, difficult, or strenuous, and repetitive

tasks. RAS has played an important role in extreme and hazardous environments that can be dynamic, uncertain, and probably even unknown, which bring many challenges to RAS

Effective communication between humans and robots requires that robots are able to communicate with socially acceptable responses and common-sense knowledge to handle a variety of situations with clear interpretation and understanding of their complicated semantics. The effective recognition of human emotions is a major challenge because human emotions are complicated and uncertain, and even single expressions presented by different individuals can be interpreted very differently.

Another key challenge for trustworthy HMI lies in Theory of Mind (ToM), i.e., the ability of humans to infer the intentions and beliefs of others [36]. To ensure trusted HMI, we should improve the understanding of the thoughts of the two communicating entities. The robot's understanding of the human ToM in terms of intention, knowledge, and competence can improve the quality of trustworthy interaction [37]. Currently, some computational ToM models have been proposed to improve the robot's understanding of the human user's intentions and trustworthiness [38][39].

The inability to exhaustively test a complicated HMI system brings the most critical challenge to system verification. This may lead users to face many unexpected situations in use of these RAS that are tested incompletely. Therefore, RAS should have the ability to learn, adapt, and respond to unforeseen circumstances in a dynamic and changing world. Implementing such a capability of RAS will be an important step to move forward in the future.

E. Challenges of Implementing RAS Ethics

Ethics is a study of the moral principles that govern a person's behaviour or the conduct of an activity. It involves systematising, defending, and recommending concepts of right and wrong behaviour [40]. RAS ethics is intended to address how human developers, manufacturers, and operators behave to minimise the ethical harm that RAS might cause due to unethical design or misplaced applications. Professional bodies have actively begun developing recommendations and policy statements. For example, IEEE has published the document "Ethically Aligned Design" to promote public understanding of the importance of taking ethical considerations into account in the development of autonomous and intelligent systems [41], and the European Group on Ethics in Science and New Technologies published a call for a "Shared Ethical Framework for Artificial Intelligence, Robotics and Autonomous Systems" in March 2018 [42].

Researchers from the UK-RAS network have identified seven ethical issues in RAS, such as Bias, deception, employment, opacity, safety, oversight, and privacy [43], of which safety and privacy have been addressed in the two properties of safety and security of RAS in this research. Bossmann [44] identified top nine ethical concerns related to AI, including unemployment, inequality, humanity, Artificial Stupidity, Racist Robots, security, Evil Genius, Singularity, and Robot Rights. The European Parliamentary Research Service (EPRS) divided AI ethics into three phases of concerns: immediate, current concerns (e.g., privacy and bias), near- and-medium term concerns (e.g., impact of AI and robots on jobs and workplaces), and longer-term concerns (e.g., the possibility of superintelligence) [45]. Therefore, the ethics challenges of RAS can be summarised in three phases on a timeline.

(1) *The immediate concern* is to bring ethics into the design of an autonomous system. The first key challenge is the bias of designers, manufacturers, operators, and especially ML algorithms. Gerdes and Thornton [46] studied the implementation of ethics in autonomous vehicles, relating ethics to constraints or costs in the design. An example key challenge is dealing with the trolley problem. Furthermore, human biases, e.g., in the selection of learning data, can lead to ethnic biases in the results of automated selection procedures. The second key challenge lies in deception. Boden et al. [47] stated that robots are manufactured artifacts that should not be deceptively designed to exploit vulnerable users. The third key challenge is to make decision-making transparent on RAS to enable control and avoid oversight. The fourth key challenge is to create regulations and laws that are accepted and applied by the public to shape the behaviour of designers, manufacturers, and operators and to ensure the implementation of ethics.

(2) *Near- and medium-term concerns* are about the roles of robots in society. With the deployment of RAS, many jobs of humans could be taken over by robots. Humans will face the challenge of how to create a room to assume advanced roles for humans and how to shape the hierarchy of the workforce in society. Robots' rights are another challenge. The fact that the robot Sophia became a citizen of Saudi Arabia in 2017 attracted a lot of public attention. It was the first humanoid AI robot in the world to become a citizen of a country [48]. However, there are no regulations or laws that clearly articulate the rights of robots yet.

(3) *The longer-term* concern is about the possibility of robots reaching or surpassing human capabilities (so-called superintelligence) [45]. The challenge is to govern the innovation of RAS and avoid superintelligence and singularity where technological growth becomes uncontrollable and irreversible, leading to unpredictable changes in human civilization.

These three phases of concerns require human-centric AI for the development of a trustworthy RAS, which is discussed in Section V.

IV. Opportunities of AI Technologies

AI technologies are used in many fields, such as extreme environments, social and health care, manufacturing, and military. For example, to ensure the safety and reliability of robots with a high degree of operational autonomy under uncertain conditions, Zhao et al. [49] developed a Bayesian inference and uncertainty model based on a layered Markov model, which was verified using the example of unmanned underwater vehicles in extreme environments. The research by Zhou and Yang [50] shows that a deep convolutional neural network for 2D biomedical semantic segmentation outperforms conventional methods in terms of both accuracy and degree of automation. Lee et al. [51] investigated the state of AI

technologies and their power in ecosystems to implement the requirements of Industry 4.0. In addition, AI-based trajectory and payload optimization of rovers are employed for the Mars mission of NASA [52].

The success of AI applications for various purposes has shown that AI plays a key role in implementing trustworthy RAS. However, much critical work remains to be done in most application domains. For example, self-driving vehicles have not yet reached the full level of autonomy. Many lessons learned from the accidents of safety critical systems (e.g., vehicles on roads or airplanes in sky) tell us that, to improve the trustworthiness of RAS, trust properties must be added to the performance and quality of services provided by RAS. Namely, RAS should be robust to all system problems, safe for any uncertain and dynamic environments, secure to all cyber threats, cyber-attacks and cybercrimes, and tolerant to all user mishandling and allow users to intervene at any execution point even if it is a fully autonomous system.

A. AI for RAS Safety

A monitoring system is essential to ensure the safety of RAS [1]. NASA updated a conventional monitoring system architecture by including user inputs in the monitoring loop, as shown in Fig. 2.

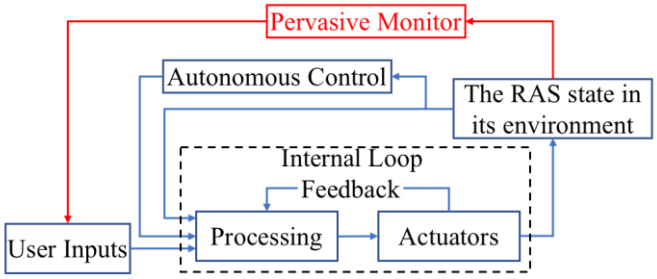

Fig. 2. Pervasive Monitoring Architecture, derived from [1]

Safety-critical systems, including various types of RAS, such as UAVs, aircraft, autonomous vehicles, use a Safety Instrumented System (SIS) with specific control functions to maintain safe operations of their processes when unacceptable or hazardous conditions occur. Many techniques can be used to implement SIS for different types of autonomous systems. A sensor system should have the functions of data acquisition, data pre-processing, and perception, etc. A key function of SIS is to detect anomalies in its environment. Advanced sensing techniques are essential to ensure the performance of SIS. The most commonly used sensors include laser sensors (LIDAR) [53], visual sensors [54], radar, GPS, infrared sensors [55] and ultrasonic sensors [56]. Aviation systems have the highest safety standard for all autonomous systems. In an aircraft, there are many monitoring subsystems, such as instrument monitoring, system monitoring, and environmental monitoring. Autonomous navigation is an important challenge in RAS. It is to enable a RAS to navigate safely and autonomously in an environment that can be uncertain and dynamic. For example, the capability of UAV navigation depends on advanced sensor systems and intelligent control algorithms that refer to the operational status and spatial information of all surveillance systems and adjust flight behaviour accordingly.

Various AI techniques, especially ML, have been used to perceive the environment and control the navigation of RAS. For example, Ouarda [57] proposed a neural path planning approach for mobile robots; He et al. [58] developed a linguistic decision tree model to solve the classical robot routing problem by decomposing a robot's task or behaviour into a few atomic units; Huang et al. [59] developed a dynamic obstacle detection system with a support vector machine using the space-time feature vector of LIDAR for driverless cars; Wu et al. [60] transformed a path planning task into an environment classification task, for which a Deep Convolutional Neural Network (DCNN) was used to determine the direction of robots; Zhu et al. [61] proposed a two-stage speed sign detection system for autonomous vehicles in dynamic environments based on a DCNN for salient target detection based on the Markov chain for background absorption; Mohanta and Keshari [62] developed a two-stage path planning using a probabilistic roadmap method to determine the shortest path between the start position and the target position in a pre-known cluttered environment, adjusting the head angle to ensure smooth turns along then tirepath. In the learning process, an autonomous vehicle should have the ability to detect any anomaly case in the environment, and record the experience, thus to produce an adaptive learning process.

However, there still lacks of an effective mechanism or architecture for implementing efficient online ML algorithms, which can adapt to a dynamic environment. Also, improving perception (e.g., obstacle detection), positioning accuracy, decision accuracy, and environmental resilience are still major challenges for RAS. Reinforcement learning could provide a technique to improve perception and decision during the RAS learning process by learning an optimal or nearly-optimal policy that maximises the "reward function" with respect to the dynamic environment and target. Fig. 3 illustrates the roadmap to improve the safety of RAS in which the four types of AI technologies play an important role such as pattern recognition for obstacle detection, autonomous control, anomaly detection and adaptive learning, and reinforcement learning. They are commonly used to overcome the challenges caused by dynamic environment and uncertainty, real-time performance and fast response, unknown errors, and lack of prior knowledge.

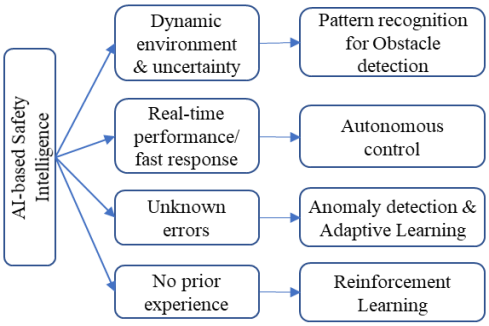

Fig. 3 AI techniques for the four challenges of RAS safety.

B. AI for RAS Security

RAS operates in real time. Human power is often not quick enough to protect RAS. Nevertheless, security automation may be necessary to effectively and efficiently protect RAS, significantly mitigate the risk of cyber threats, and enable prompt response to cyber incidents caused by cyber-attacks and

cybercrimes, reducing the impact of cyber incidents on RAS and economic losses. AI technology is essential for creating security automation for RAS. Unlike other problem domains, RAS cybersecurity must withstand the malicious behaviour of determined and sophisticated attackers, as required for all IoT enabled CPS [63]. RAS is usually connected to the Internet to enhance its computing capacity, storage capacity, accessibility, usability, and flexibility, etc., but could be a target of hackers for various reasons. Cyber Intelligence should be able to secure the benefits of the Internet connected world for all. An architecture for RAS with security automation is required to enable the implementation of "security by design" required by Industry 4.0 while maintaining adaptive, self-learning and autonomous security [64].

AI techniques have been applied to cybersecurity in two areas: In access control, pattern recognition techniques have demonstrated their performance in biometric authentication (e.g., fingerprint, face, iris, and palm) and signature and keystroke verification. For example, Fang et al. [65] proposed a fast and holistic authentication and authorization approach to analyse the complex dynamic environment through online ML and trust management to achieve adaptive access control. User Entity Behaviour Analytics (UEBA) is a new security process that uses ML algorithms and statistical analysis to detect network attacks in real time [66].

ML techniques have played an important role in data-driven cybersecurity because they bring two key benefits to threat intelligence: first, machines can handle huge amounts of data and their complex relationships, which it is impossible to be done by humans; second, machines can implement cybersecurity automation, which is impossible for humans. Much research has been done in this area, e.g., intrusion detection [67], anomaly identification [68], web robot detection [69], and malware detection [70].

The goal of ML-based intrusion detection systems (IDS) is to improve the accuracy and reduce the false alarm rate of unknown attack detection. Since attack techniques are constantly evolving, IDS should firstly work passively at the network level to detect intrusions to reduce the impact of cyber-attacks on RAS; secondly, IDS should be adaptive to detect new attacks. To secure RAS enabled by IoRT, intrusion detection is required in edges. However, the limited computational capacity and critical real-time performance requirements of edge devices may limit the capacity of edge IDS. Therefore, developing effective and efficient IDS is a critical challenge. ML algorithms have been developed to detect DoS attacks and secure IoT-enabled systems [71]. With the growing number of cyber-attacks and increasingly complex IT environments, an intelligent incident response mechanism is more than just a set of instructions. Automation is the best approach to enabling rapid incident response [72]. An effective and efficient threat analysis based on a socio-technical model with alert enrichment and prioritisation of actions on RAS could be a viable solution for securing IoRT-enabled systems.

However, as mentioned in [73], ML models, as the main driver of cognitive cybersecurity, may be hacked, since the implementation of ML algorithms is a programme or a function in a programme. Hackers could not only change the code, but also insert training samples or replace them with adversarial samples, which is an instance with small, intentional feature perturbations that cause a ML model to make an incorrect prediction. Adversarial examples make ML models vulnerable to attacks. The consequences of an adversarial example could be severe for RAS and directly damage the trustworthiness of RAS. For example, a driverless car could crash into another car or into pedestrians, because it ignores a stop sign if a hacker placed an image over the stop sign.

Therefore, the protection of ML models should address two challenges: First, the code of the ML models must be secured, which includes the training code and the testing code, and second, the adversary-fitting problem must be solved. For the first challenge, the effort is the same as for protecting general software systems; in real applications, it is difficult to verify whether the collected data is false or not. The classical adversarial defence techniques described in [74] include:

(1) Adversarial training, an intuitive defence method against adversarial samples that attempts to improve the robustness of a neural network by training with adversarial samples.
(2) Randomisation schemes to mitigate the effects of adversarial perturbations in the input/feature domain on Deep Neural Networks (DNNs) that are robust to random perturbations.
(3) Denoising (e.g. GAN-based input clean-up) is a method for mitigating adversarial perturbations/effects due to attackers. There are two ways of denoising: (i) partially or completely removing the adversarial perturbations from the inputs, and (ii) mitigating the effects of adversarial perturbations on the high-level features learned by DNNs.
(4) Provable defence techniques based on well-defined attack types.

ML as a key technology is driving the development of security automation, but we need to ensure that correct (or trustworthy) features or data are fed into ML models for their own security. Therefore, the robustness and security of data collection systems and ML models must be investigated during system design. Fig.4 shows the AI roadmap for implementing RAS cybersecurity automation.

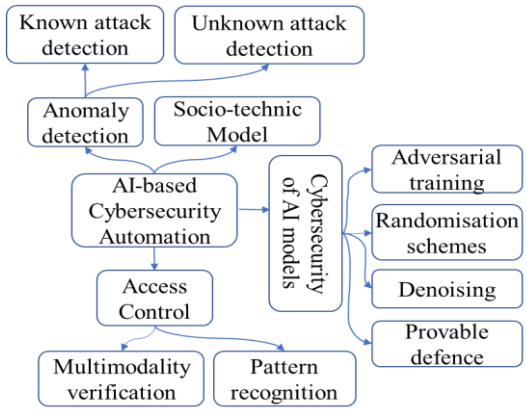

Fig. 4 The roadmap of RAS cybersecurity automation.

Regardless of the type of ML technique used to detect cyber-attacks or anomalies, key performance indicators include accuracy, F-measure, Confusion Matrix, ROC curves, Mean-

Squared Error, Standard Deviation, etc. While the goal in anomaly detection is to improve the true-positive rate, the false-positive rate cannot be ignored. Real-time performance is imperative, especially for applications of RAS, even though the computational capacity of RAS may be limited for on-board countermeasures to secure RAS. This presents critical challenges in implementing security automation.

C. AI for the Health of RAS

One of the most important factors that affect the reliability of a system is the state of the system and its ability to self-diagnose. Fault diagnosis is about identifying a faulty system by observing its behaviour. It can be an important technique to ensure the safety of RAS. With the development of AI techniques, many new methods have been applied for fault diagnosis. Dynamic artificial immune system is one of the AI methods with strong self-learning and self-adaptation capabilities.

Since 1980s, analytical redundancy methods have been a main trend for fault diagnosis [75]. Fig. 5 illustrates the framework of an analytical fault diagnosis model, where fa is an actuator fault, fc is a process/component fault, and fs is a sensor fault.

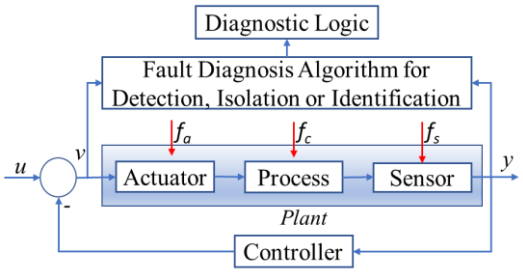

Fig. 5. Analytical Fault Diagnosis [15]

A fault diagnosis algorithm is performed to check the consistency of the feature information of the real-time process carried by an input u and the output y against the prior-knowledge of a healthy system, and a diagnosis decision is then made by the diagnosis logic. With the development of AI or ML techniques, diagnostic algorithms can be implemented using a ML model that can be trained with a set of historical data from the status information of sensors, actuators, and components, including fault information, f_a, f_c, and f_s.

Fault diagnosis usually has three tasks, namely fault detection, fault isolation and fault identification. As the most basic task of fault diagnosis, fault detection is used to check whether there is a malfunction or a fault in the system and to determine the time of occurrence of the fault, and it can be transferred to a decision-making problem in ML. The aim of fault isolation is to determine the location of the faulty component, and it could be transferred to an optimization problem. Fault identification is to differentiate the classes, levels and dimensions of a fault, and can be solved as a classification problem using ML techniques. From a technical point of view, fault diagnosis can be divided into four categories: model-based fault diagnosis, signal-based fault diagnosis, knowledge-based and hybrid methods.

Model-based diagnosis is concerned with locating faulty components of a system based solely on its structure and behaviour. It performs various diagnostic tasks by online reasoning and inference of the global behaviour of a system from the automatic combination of local models of its components. Clearly, ML techniques can be used to build models that represent the relationships between inputs and outputs (Eq. (1)), where v is the sum of the input u and the feedback from the controller in Fig. 3.

$$\hat{y} = \mathcal{L}(f(v)). \quad (1)$$

The system checks whether the output y is consistent with the model output ŷ; if they are inconsistent, the system is considered faulty. For example, Hashimoto et al [76] developed a multi-model approach to identifying three failure modes (e.g., hard failure, noise failure, and scale failure) of faults in the five internal sensors on a mobile robot based on Kalman filters.

Instead of explicit input-output models, signal-based fault diagnosis usually uses three types of measured signals: time domain, frequency domain, and time-frequency domain to determine faults [15]. Since the measured signals reflect faults in the process, we can extract the features from the measured signals and use the symptom analysis and prior knowledge about the symptoms of the healthy systems to make a diagnosis decision. To implement automatic fault diagnosis, various ML techniques can be used with the measured signals, the features extracted from the signals, or the raw sensor data. For example, Eski et al. [77] used artificial neural networks to predict faults based on the noise and vibration of robot joints; Cho et al. [78] used neural networks to estimate the fault torque of an adaptive actuator for robot manipulators; Ran et al. [79] categorised three types of fault diagnosis and prognosis in predictive maintenance systems, such as knowledge-based, traditional ML-based, and Deep Learning-based approaches.

Knowledge-based fault diagnosis methods can effectively use expert knowledge and experience to make judgments [80]. Knowledge representation provides clues for ontology reasoning in fault diagnosis. However, in the actual fault diagnosis process, it may be difficult to determine the relationship between the component fault phenomenon in RAS and the fault cause. A fault phenomenon on a component may have many causes, while a fault may show different types of phenomena on different components. Due to the complex structure of RAS, the error with multiple causes and the suddenness of an error, the combination of empirical knowledge and mechanism principles can be used to solve various fault problems. While fuzzy knowledge representation provides an effective approach to representing the uncertainty of knowledge [81], a dynamic uncertain causality graph can be useful to illustrate the relationship between fault phenomenon and causes [82].

To improve the accuracy of fault diagnosis for RAS, a hybrid approach combining different ML models and prior knowledge in the problem space can be a solution. For example, Nadeer et al [83] developed an online fault diagnosis method for a spark ignition vehicle engine using a hybrid model with three stages: a single extended Kalman Filter estimator, a residual prediction stage, and a fault detection and isolation stage.

Diagnostic accuracy, real-time performance, and data availability are key challenges in data-driven fault diagnosis,

just as they are in data-driven cybersecurity. Effectively and efficiently assigning a root cause to a fault is a critical challenge for a complex system with many correlated components. Although AI optimization techniques can improve the accuracy of fault allocation, it might be difficult to achieve real-time performance. Fig. 6 illustrates the roadmap of AI techniques in edges for improving the health diagnosis capacity of RAS, where the four methods are supported by AI techniques and underpinned by advanced sensor technology and human intervention to support the implementation of the three tasks of fault diagnosis.

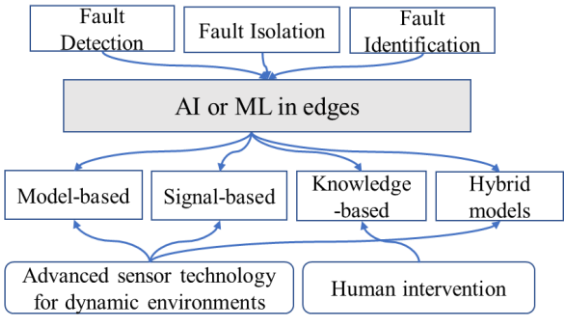

Fig. 6 Roadmap of AI-based RAS fault diagnosis.

Multiple homogeneous or heterogeneous robots can be connected together to form a robot swarm. Since the individual robots have integrated processing, communication, and sensing capabilities, they can interact with each other and respond autonomously to their environment. A swarm robot system uses swarm intelligence for a mission in extreme and hazardous environments or for entertainment. Such a large swarm system without human intervention requires the use of autonomous self-diagnosis, self-healing, and self-replication under certain circumstances that humans cannot perform efficiently. For example, Dai et al [84] proposed a multifunction model for a swarm robot system that includes virtual neurons running in a robot as a background process for perception and reflexes, autonomous self-diagnosis through consequence-based prescription, autonomous self-healing, and self-replication. A swarm intelligence could help for the implementation of such self-functions performed by correlated autonomous robots in a swarm without human intervention.

D. AI for Trusted Human-Machine Interaction

Trusted Human-Machine Interaction (HMI) is a challenge for human-centred Artificial Intelligence (HAI) [85]. HMI requires the harmonious collaboration of interdisciplinary fields. It involves human behaviour and mind modelling to improve robotic recognition, knowledge acquiring, representing, and manipulating at the human-level of reasoning and decision making; thus, eventually instantiating physical actions that are legible to and coordinated with humans. HMI researchers strive to leverage advanced technologies from the fields of AI and quantum computing into user-friendly human-machine interaction systems that are oriented toward our lives. Fig. 7 provides the AI roadmap in Human-Robot Interaction.

Trusted HMI can benefit from cognitive and developmental robotics models. The developmental studies on ToM in children [86] have informed the design of developmental cognitive architectures for artificial ToM in robots [87][88]. The cognitive model of ToM in [87], built on an operationalisation of psychology experiments on children in [86], allows a Pepper robot to predict the trusted/untrusted behaviour of people and follow, or reject people's recommendations. Developmental studies of ToM can also be used to investigate the trust in child-robot interaction scenarios [89].

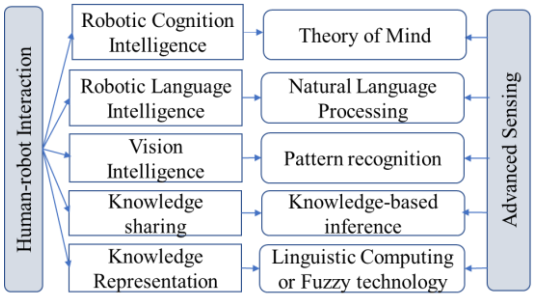

Fig. 7 AI roadmap in Human-Robot Interaction.

Natural Language Processing (NLP) is an important technique for improving the cognitive capacity of HMI, and the NPL capacity of RAS is an important indicator of human-level machine intelligence [81]. For decades, classical shallow ML models (e.g., Support Vector Machine and logistic regression) have been used to solve NLP problems with high-dimensional and sparse features. In recent years, deep neural networks based on dense vector representations have shown superior performance for various NLP tasks [90]. NLP mainly includes two types of tasks: natural language understanding (NLU) and natural language generation (NLG). NLU includes the tasks of mapping the given input in natural language into useful representations, such as rule-based machine translation and analysis of various aspects of the language. NLG includes text planning, sentence planning, and text implementation.

Robot Vision uses a combination of camera hardware and computational algorithms to enable robots to process visual data from the world. Visual understanding capabilities are added to a robot so that the robot can perceive human nonverbal behaviour and interact naturally with humans through body gestures, facial expressions, and postures. For example, an intelligent robot has been developed to assist physicians in surgical procedures using two cameras, a Near-Infra Red (NIR) camera and a panoptic 3D camera, to create high-contrast areas in the 2D NIR images [91]. Deep learning is a competitive technology for various computer vision benchmark problems such as image classification, object detection and recognition, semantic segmentation, and action recognition [92].

Tactile sensing is a key technology to ensure that physical human-robot interactions are safe when such interactions require physical contact between humans and robots in a shared workspace. A tactile sensor can be used as an artificial sensitive skin of a robot, providing not only safety-related functions but also touch-based motion control of the robot, thus improving human-robot interaction [93]. Distributed tactile sensors can be easily deployed at different locations on the robot body. Their ability to estimate contact forces and create a tactile map with accurate spatial resolution enables the robot to safely handle intentional touches and avoid unintentional collisions in safe

human-robot collaboration tasks based on multi-sensor fusion [94].

Knowledge extraction and sharing between humans and machines is crucial to the dynamic process of human-machine interaction. The knowledge from the input of the user and the output of the human-machine interface can be extracted by the machine. Artificial intelligence has made some progress in how the robot recognises and understands the external information and then performs the appropriate action. Various reasoning systems have been used for knowledge-based inference in human-machine interaction [95]. Tran et al. [96] have shown that layer-wise extraction can improve the performance of Deep Belief Networks, and they have proposed a symbolic characterization approach to inserting prior knowledge and training Deep Networks.

Knowledge representation can support the transformation of knowledge into the user interface environment by modelling the abstractions of knowledge [97]. It is important to create an adaptive knowledge representation process for HMI automation [98]. Devlin et al. [99] designed a model of Bidirectional Encoder Representations from Transformers (BERT) based on a deep bidirectional architecture to pre-train deep bidirectional representations from unlabelled text that successfully support various NLP tasks. Zadeh stated that fuzzy techniques can provide an effective approach to represent the imprecision and uncertainty of knowledge; fuzzy logic is a precise logic of imprecision and approximate reasoning, and a fuzzy set is a class with fuzzy boundaries [81]. Therefore, recent developments in fuzzy technology can help make a system more robust.

One of the critical properties of the future (cognitive and developmental) robots is that they can update their internal representations, knowledge, and functions. This type of property was not observed in conventional systems. To overcome this problem, we should develop a human-robot interface for facilitating people to understand the cognitive states of robots, i.e., dynamic internal representations. For example, Hafi et al. [100] developed a mixed reality interface to enable users to intuitively visualise the current state of the robot perception and naturally interact with it.

Explainable AI is significantly important for implementing trusted human-machine interaction, because it will help people understand autonomous adaptive intelligence. There has been much research that focused on transparency in decision making. For example, He and Lawry developed a linguistic attribute hierarchy [101], and a linguistic CMAC neural network [102] for decision making. He et al [58] applied a Linguistic Decision Tree to solve robot route learning by establishing a transparent relationship between robot behaviour and environmental changes. Recently, Grossberg [103] explored the path towards explainable AI and Autonomous Adaptive Intelligence, by examining six facets of interpretable autonomous intelligence: (1) solution (e.g., ARTMAP, a type of self-organizing neural network architecture [104]) to solve the black-box problem of neural networks (e.g., BP networks or deep neural networks); (2) rule-based reasoning (e.g., fuzzy ARTMAP) to explain decision or classification; (3) explainable visual and auditory perceptions; (4) explainable emotions during cognitive emotional interactions; (5) explainable representations of robot behaviour; (6) explainable autonomous adaptive intelligence.

V. HUMAN-CENTRED AI FOR TRAS

There are many different concerns about human-centred AI. In the context of TRAS, human-centred AI is a means of improving the relationship between humans and RAS. Fridman [105] believes that the implementation of autonomous vehicles is the problem of integrating human-machine interface, machine intelligence, psychology, and policy, and he proposed seven principles for the practice of human-centred autonomous vehicle: (1) sharing autonomy with humans, (2) learning from data, (3) human perception, (4) sharing perceptual control, (5) deep personalisation, (6) imperfect by design, and (7) system-level experience. Nevertheless, human-centric AI for TRAS implementation may need to consider the following aspects.

A. Acceptance Model of TRAS

Human acceptance of RAS determines the requirements for human-centric AI that enables the implementation of TRAS. We propose an acceptance model for TRAS that includes two aspects: worthiness and trustiness, as shown in Fig. 8.

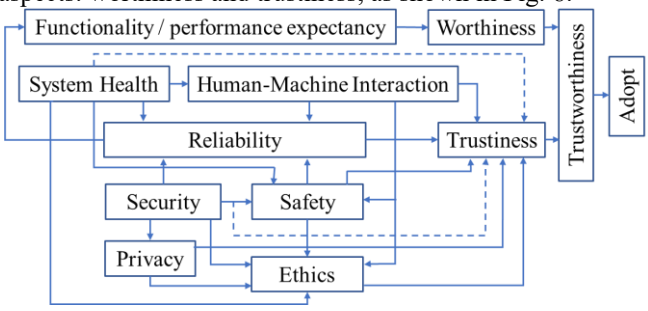

Fig. 8. Acceptance model for trustworthy RAS.

Worthiness represents the quality of being good enough or suitable, trustiness represents the quality of being loyal or reliable, and acceptability is the action or fact of choosing to adopt, follow, or use something. Only when both worthiness and trustiness are integrated in RAS, people would accept and adopt it. The functionality/performance of RAS represents the worthiness of RAS, while all five properties and the reliability of RAS represent the trustiness of RAS. The four properties in safety, security, system health and human-machine interaction are directly related to the reliability of RAS.

Many models have been developed to explain human behaviour and acceptance of new technologies. People may not be aware of the importance of cybersecurity. Therefore, Zhang et al. [106] did not include security as an important element in the acceptance model of Connected and Autonomous Vehicles. He et al. [107] argued that security can directly affect reliability and privacy, as evidenced by a Denial of Service (DoS) attack, where a flood of random packets is sent to a target system and can cause the system to malfunction, resulting in, for example, a stuck or inoperable throttle, raising potential safety concerns [108]. Therefore, cybersecurity is directly related to safety, but indirectly related to trustiness. Privacy depends on cyber security protection, but is directly related to the RAS trustiness.

The health state of RAS could affect its interaction to its environments. An anomaly state could cause malfunctions of

the RAS, which result in wrong behaviour of the RAS, even producing wrong interaction to its environment, and thereby threaten the safety of the RAS. Also, fault tolerance should reflect ethics requirements. Therefore, the health of RAS is indirectly related to trustiness.

The reliability of RAS could affect the functionality and performance of RAS, which is directly related to worthiness. Reliability can be defined as the likelihood that a system will produce correct results when in use. Users expect a system to be highly reliable when they choose to use it. Therefore, RAS reliability is directly related to trustworthiness. A reliable system does not silently continue and deliver behaviour that may be incorrect. The reliability of a system can be improved by features that help prevent, detect, and tolerate errors that might be caused by problems related to safety, security, health and human-machine interaction. These four properties are thus directly linked to reliability.

Ethics is a key property directly related to the trustiness of RAS. As mentioned in Section III, ethical issues include safety, security, privacy, and human-machine interaction. Therefore, these four properties are directly related to ethics as well.

The human-machine interface can have a direct impact on safety. The dynamics of human-machine interface technology requires the development of innovative human-machine interface approaches and methods to support the design of complex socio-technical systems within the framework of existing ethical and legal regulations [109]. Therefore, the human-machine interface is directly linked to ethics.

In Fig. 8, the solid lines represent direct relationships and the dashed lines represent indirect relationships. It is obvious that the proposed acceptance model is applicable to autonomous vehicles. Trustworthy autonomous vehicles that meet the requirements under the acceptance model would be accepted. In different application domains, the relationships between the blocks in the model may be weighted differently. For example, household robots may have fewer safety requirements for trustworthiness than an autonomous vehicle on the road.

B. Towards Human Level Intelligence

In line with PwC [110], we can briefly divide robotic intelligence into three levels: assisted intelligence, where AI replaces many of the repetitive and standardized tasks performed by humans (e.g., manufacturing machines); augmented intelligence, where humans and machines learn from each other and redefine the scope and depth of their joint work (e.g., surgical robots); and autonomous intelligence, where adaptive and continuous machines take over the entire process of perception, decision-making, and action, independent of human intervention (e.g., fully autonomous vehicles). Autonomous Intelligence is the most advanced robotic intelligence. However, from a human perspective, the autonomy of machines should be limited, and at least humans can take control of them at any point of execution.

As Zadeh states in [81], humans have many remarkable capabilities, two of which are particularly important: (1) the capability to reason, converse and make rational decisions in an environment of imprecision, uncertainty, incompleteness of information, partiality of truth, and possibility; (2) the capability to perform a variety of physical and mental tasks without measurement or computation. There is ample evidence that machine intelligence has made some progress in the first capacity, in terms of linguistic intelligence, visual intelligence, auditory intelligence, tactile intelligence, spatial intelligence, and emotional intelligence (e.g., the humanoid and programmable robot Nao [111]). This shows that the cognitive abilities of a robot have made some progress. Although robotic cognitive capabilities are constantly evolving, we are still far from the second capability, which requires a machine brain capable of perceiving complex environments and making decisions quickly. With the first capability, robots are able to perceive simple environments or events, how such simple single perceptions can be comprehensively and efficiently integrated to form an embryonic, responsive nervous system for a robot is still unclear.

Clearly, achieving human-level machine intelligence is still an elusive challenge. As we move towards a fully autonomous system, machine intelligence is also approaching human intelligence. The dimensions of the environments are now increasing as cyberspace is added to the physical execution environments inside and outside the systems. A prerequisite for achieving human-level machine intelligence is the mechanisation of these capabilities [81]. In other words, a prerequisite for implementing TRAS is to automate the four capabilities of RAS, which are consistent with automating the functionality of RAS with respect to different application domains. Creating a general human-level intelligence by integrating the automation of all the functions with the respect to the five properties of RAS is not only a critical challenge, but also a formidable work that requires many more generations of research and innovation. The human acceptance model for RAS can provide a framework for the requirements of a trustworthy RAS with human intelligence.

C. Augmented Human Capabilities

An important goal of developing human-centred AI to implement TRAS is to enhance human capabilities and skills. RAS should communicate and collaborate effectively with humans, each partner bringing its own superior capabilities to the partnership - rather than looking for AI to replace humans, using AI to complement humans and human intelligence. In this way, intelligent systems that put humans at the centre and intelligent human-machine interfaces could either augment existing human capabilities and skills or create new ones. Such augmented humans and intelligences will allow humans to go beyond their current capabilities and have new experiences [112]. Automated machines have long operated in factories to perform highly repetitive and physically demanding tasks more efficiently and productively. This is referred to as "automated" intelligence, of which the automotive industry provides a good example. The report by Oxford Economics [113] predicts that 6.6 million jobs could be lost in ASEAN-6 economies by 2028 as a result of the introduction of new technologies. This will require a redesign of entire business processes and a redefinition of the jobs people do, similar to how the job of bank clerk was redefined with the advent of ATMs.

D. Focus on AI's Impact on Humans

Technology is trusted if it benefits humans, is demonstrably fair, safe and reliable, is well regulated, and can be investigated if errors occur. Obviously, RAS has not yet reached this stage. Trust and justified confidence can accelerate technology adoption and job creation and prevent a backlash against RAS, but only if trust is not misplaced [9].

Robots can take over many human jobs. However, this does not mean that robots can replace and control humans. To develop trustworthy RAS, we need to understand how artificial intelligence behaves in practise and how it affects humans. This requires humans to be careful and follow rules when developing fully autonomous systems to avoid any uncontrollable robotic intelligence. In the development of RAS, we should find a solution that allows the human to interrupt the work of the robot at any point of execution and limit the situation to one that does not harm the human. This ensures that the human can control the robots and that there is no risk of the robots harming the humans. This is in line with the first principle of sharing autonomy with humans in the loop proposed in [105].

Many existing or yet-to-be-developed AI technologies can be used for both commercial and military applications. The greatest threat posed by AI is potential weaponization [114]. In contrast, we should encourage the development of AI-enabled RAS in areas where RAS could (1) help people work in extreme and dangerous environments, (2) improve the capacity of health and social care, (3) increase the ability to manufacture and produce food, (4) reduce damage to the Earth, (5) help recover the damage we have done to the planet, and (6) explore space. All of these and other global challenges should give rise to AI-powered, mission-based innovation and bring together citizens, scientists, and engineers to address them. The AI for Good Global Summit in June 2017 discussed how AI could chart a development course that can help achieve the goals of United Nations' Sustainable Development Goals [115]. An AI partnership composed of more than 100 industries, including major companies such as Google, Microsoft, and IBM, was formed to bring diverse, global voices together to realise the promise of AI [116].

Nevertheless, adherence to Asimov's three laws for robots [117] is the essential requirement. Murphy and Woods [118] provide three alternative laws for responsible robotics, of which, the most important is that “A human may not deploy a robot without the human–robot work system meeting the highest legal and professional standards of safety and ethics”. These laws should be considered in the design of a trustworthy RAS equipped with human-centred AI.

VI. Conclusions

This research provides a comprehensive view on the concept of human-centred AI for trustworthy RAS with respect to safety, security, health, human-machine interaction and ethics. Fig. 9 concludes the concept of human-centred AI that underpins the goal of a fully trustworthy RAS. The main contribution of the research is that, for the first time, the five key facets of human-centred AI (HAI) for trustworthy RAS have been systematically explored in terms of their challenges in implementing trustworthy RAS, the role and the roadmap of AI in implementing the five facets of a trustworthy RAS, thus providing a structural approach to designing trustworthy RAS.

While the performance and functionality of RAS are important aspects of RAS, the five properties discussed ensure the trustworthiness of RAS. A new acceptance model for RAS is proposed to ensure both worthiness and trustiness of RAS for human acceptance and adoption of RAS. It provides a framework of requirements for implementing trustworthy RAS with human-centred AI. The new concept of human-centred AI (HAI) promotes collaboration between technological innovation and humanistic and ethical considerations with three goals: (1) advancing the technical frontiers of human intelligence; (2) expanding human capabilities; and (3) focusing on the positive impact of AI on humanity. Finally, we emphasise that a trustworthy RAS must at least obey Asimov's three laws of robotics.

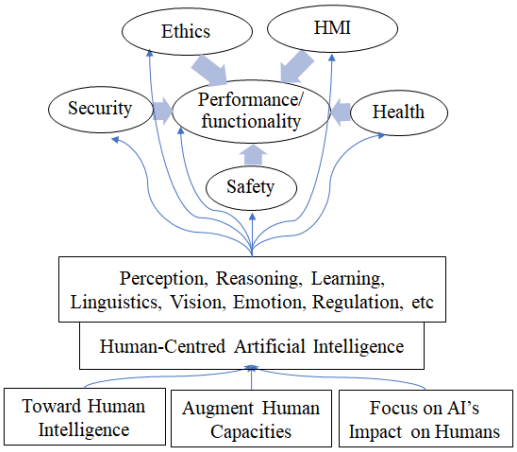

Fig. 9 Human-centred AI for Trustworthy RAS.

References

[1] E. E. Alves, D. Bhatt, B. Hall, K. Driscoll, A. Murugesan, and J. Rushby. 2018, Considerations in assuring safety of increasingly autonomous systems. NASA/CR–2018-220080, NASA, July 2018.

[2] The Internet of Robotic Things, Technology Analysis Report, ABI Research, 2014. Accessed on 21/10/2021.

[3] S. Levin and J. C. Wong, Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian, The Guardian, 19 Mar 2018. Accessed on 21/10/2021.

[4] A. Drozhzhin, Black Hat USA 2015: The full story of how that Jeep was hacked, 7 Aug. 2015. Access on 21/10/2021.

[5] BBC News, Boeing 737 Max Lion Air crash caused by series of failures, 25 October 2019. Accessed on 21/10/2021.

[6] European Group on Ethics in Science and New Technologies, 2018, Statement on Artificial Intelligence, Robotics and ‘Autonomous’ Systems, Mar. 2018. DoI: 10.2777/531856.

[7] S. Nyholm and J. Smids, The Ethics of Accident-Algorithms for Self-Driving Cars: An Applied Trolley Problem? Ethic Theory Moral Prac 19, 2016, pp. 1275–1289. DoI: 10.1007/s10677-016-9745-2

[8] T. Evas, European framework on ethical aspects of artificial intelligence, robotics and related technologies, European added value assessment, European Parliamentary Research Service, PE 654.179 – Sept 2020. DoI: 10.2861/94107.

[9] UKRI Trustworthy Autonomous Systems Programme - Town Hall meeting, 23 Sept 2019. Access on 21/10/2021.

[10] NIST, Framework for Cyber-Physical Systems: Vol. 1, Overview Version 1.0. Jun 2017. DoI:10.6028/NIST.SP.1500-201.

[11] A. Zacharaki, I. Kostavelis, A. Gasteratos and I. Dokas, Safety bounds in human robot interaction: A survey, Safety Science, 127, 2020, 104667, DoI: 10.1016/j.ssci.2020.104667.

[12] P.A. Lasota, T. Fong and J. A. Shah, A Survey of Methods for Safe Human-Robot Interaction, Foundations and Trends® in Robotics: 5(4), 2017, pp 261-349. DoI: 10.1561/2300000052.

[13] R. Subramanian, Emergent AI, Social Robots and the Law: Security, Privacy and Policy Issues, Journal of International Technology and Information Management, 26(3), 2017. Article 4.

[14] I. Kostavelis and A. Gasteratos, Robots in crisis management: A survey." International Conference on Information Systems for Crisis Response and Management in Mediterranean Countries. Springer, Cham, 2017. DoI:10.1007/978-3-319-67633-3_4.Corpus. ID: 10839315

[15] Z. Gao, C. Cecati, and S. X. Ding. A Survey of Fault Diagnosis and Fault-Tolerant Techniques—Part I: Fault Diagnosis with Model-Based and Signal-Based Approaches. IEEE Transactions on Industrial on Electronics, 62(6), Jun. 2015, pp. 3757–3767.

[16] D. A. Norman. The Design of Everyday Things. Doubleday, NY, 1988.

[17] N. Kallioinen, M. Pershina, J. Zeiser, F. N. Nezami, and G. Pipa, A. Stephan, and P. König, Moral Judgements on the Actions of Self-Driving Cars and Human Drivers in Dilemma Situations from Different Perspectives, Front. Psychol., 01 Nov. 2019. DoI: 10.3389/fpsyg.2019.02415.

[18] M. Y. Chen, Y. J. Wu and H. He*, A Comprehensive Obstacle Avoidance System of Mobile Robots Using an Adaptive Threshold Clustering and the Morphin Algorithm, The 18th Annual UK Workshop in Computational Intelligence, Nottingham Trent Uni., UK, 5-7 Sept 2018.

[19] S. Tedeschi, C. Emmanouilidis, J. Mehnen, and R. Roy, 2019. A design approach to IoT endpoint security for production machinery monitoring, Sensors. 19(10), 22 May 2019, 2355. DoI: 10.3390/s19102355.

[20] S. Katzenbeisser, I. Polian, F. Regazzoni, and M. Stöttinger. 2019. Security in autonomous systems. In IEEE European Test Symposium (ETS), Baden-Baden, Germany, 2019.

[21] N. Gisin, S. Fasel, B. Kraus, H. Zbinden, G. Ribordy, 2015. Trojan Horse attacks on Quantum Key Distribution systems, arXiv:quant-ph/0507063. DoI:10.1103/Phys RevA. 73.022320.

[22] D. D. Coleman and D. A. Westcott, Chapter 14. Wireless Attacks, Intrusion Monitoring and Policy, in book: CWNA Certified Wireless Network Administrator Study Guide. 2006, pp. 387-416.

[23] D. Antonioli, N. O. Tippenhauer, K. B. Rasmussen. The KNOB is Broken: Exploiting Low Entropy in the Encryption Key Negotiation Of Bluetooth BR/EDR, in Proc. of 28th USENIX Security Symposium, Santa Clara, CA, USA, 14-16 Aug. 2019, pp. 1047-1061.

[24] S. Rizvi, J. Willet, D. Perino, S. Marasco, C. Condo. 2017. A Threat to Vehicular Cyber Security and the Urgency for Correction Complex Adaptive Systems Conference with Theme: Engineering Cyber Physical Systems, CAS 30 Oct. – 1 Nov. 1, 2017, Chicago, Illinois, USA

[25] A. Abdulmohsan, H. He, C. Shaw and M. A. Khan, Analytical Review on Cybersecurity of Embedded Systems, IEEE ACCESS, Dec. 2020.

[26] J. A. Ambrose, R. G. Ragel, D. Jayasinghe, T. Li and S. Parameswaran, Side Channel Attacks in Embedded Systems: A Tale of Hostilities and Deterrence, Sixteenth International Symposium on Quality Electronic Design, Santa Clara, CA, 2-4 Mar. 2015.

[27] A. P. Fournaris, L. P. Fraile and O. G. Koufopavlou, Exploiting Hardware Vulnerabilities to Attack Embedded System Devices: a Survey of Potent Microarchitectural Attacks, Electronics, 6(52), 2017. DoI: 10.3390/electronics6030052.

[28] ISO/CD 10303-226 Product data representation and exchange. Application protocol Part 226, ship mechanical systems, N1015.

[29] K. Kawabata, S. Okina, T. Fujii, and H. Asama. 2003. A system for self-diagnosis of an autonomous mobile robot using an internal state sensory system: fault detection and coping with the internal condition. Advanced Robotics, 17, (2003), 925–950.

[30] G. Shi, P. Dong, H. Q. Sun, Y. Liu, and Y. X. Cheng. Adaptive control of the shifting process in automatic transmissions. International Journal of Automotive Technology, 18, 2017, 179–194.

[31] M. Yazdi, Hybrid Probabilistic Risk Assessment Using Fuzzy FTA and Fuzzy AHP in a Process Industry. J Fail. Anal. and Preven. 17, 756–764 (2017). DoI: 10.1007/s11668-017-0305-4

[32] J. Dunjó, V. Fthenakis, J. A. Vílchez and J. Arnaldos, Hazard and operability (HAZOP) analysis. A literature review, Journal of Hazardous Materials, 173 (1–3), 2010, pp. 19-32. DoI: 10.1016/j.jhazmat.2009.08.076.

[33] S. Ozturk, V. Fthenakis and S. Faulstich, Failure Modes, Effects and Criticality Analysis for Wind Turbines Considering Climatic Regions and Comparing Geared and Direct Drive Wind Turbines. Energies, 2018, 11(9), 2317. DoI: 10.3390/en11092317.

[34] R. Hegde, S. Yako, K. Post and S. Nuesch, Systems Theoretic Process Analysis for Layers of System Safety, INCOS International Symposium, 29(1), July 2019, pp. 895-909. DoI: 10.1002/j.2334-5837.2019.00642.x

[35] M. Veloso. A few issues on human-robot interaction for multiple persistent service mobile robots. In AAAI Fall Symposium Series, Arlington, Virginia, USA, 2014.

[36] S. Devin and R. Alami, An Implemented Theory of Mind to Improve Human-Robot Shared Plans Execution. The 11th ACM/IEEE International Conference on Human Robot Interaction, Mar 2016, Christchurch, New Zealand. pp. 319-326.

[37] W. Mou, M. Ruocco and D. Zanatto, A. Cangelosi. When would you trust a robot? A study on trust and theory of mind in human-robot interactions. Proceedings of RO-MAN2020, 29th IEEE International Conference on Robot and Human Interactive Communication, 31 Aug.-4 Sept. 2020.

[38] M. Patacchiola and A. Cangelosi, A developmental cognitive architecture for trust and theory of mind in humanoid robots. IEEE Transactions on Cybernetics. 28 July 2020. pp. 1-13.

[39] S. Vinanzi, M. Patacchiola, A. Chella and A. Cangelosi.. Would a robot trust you? Developmental robotics model of trust and theory of mind. Philosophical Transactions of the Royal Society B., 11 Mar. 2019. DoI: 10.1098/rstb.2018.0032.

[40] C. B. Wrenn, The Internet Encyclopedia of Philosophy, "Ethics", ISSN 2161-0002. Accessed on 21/10/2021.

[41] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition. IEEE, 2019. Access on 16/12/2020.

[42] Statement on artificial intelligence, robotics and 'autonomous' systems, Brussels, 9 March 2018. European Group on Ethics in Science and Technology, March 2018, DoI: 10.2777/531856.

[43] UK-RAS Network: Ethical Issues for Robotics and Autonomous Systems, UK-RAS white paper, 2019, ISSN 2516-5011. 18/10/2021.

[44] J. Bossmann, Top 9 ethical issues in artificial intelligence, The World Economic Forum, 21 Oct. 2016. Accessed on 21/10/2021.

[45] J. Fox-Skelly, E. Bird, N. Jenner, A. Winfield, E. Weitkamp and R. Larbey, The Ethics of Artificial Intelligence: Issues and Initiatives, European Parliamentary Research Service (EPRS), 27 Apr. 2020. DoI: 10.2861/6644.

[46] J. Gerdes and S. Thornton, Implementable Ethics for Autonomous Vehicles. In: Maurer M., Gerdes J., Lenz B., Winner H. (eds) Autonomes Fahren, 2015, Springer Vieweg, Berlin, Heidelberg. DoI: 10.1007/978-3-662-45854-9_5

[47] M. Boden, J. Bryson, D. Caldwell, K. Dautenhahn, L. Edwards, S. Kember, P. Newman, V. Parry, G. Pegman, T. Rodden, T. Sorrell, M. Wallis, B. Whitby and A. Winfield, Principles of robotics: regulating robots in the real world, Connection Science, 29(2), 2017, pp. 124-129, DoI: 10.1080/09540091.2016.1271400.

[48] C. Weller, "Meet the first-ever robot citizen — a humanoid named Sophia that once said it would 'destroy humans'". Business Insider. 27 Oct. 2017. Accessed on 21/10/2021.

[49] X. Zhao, V. Robu, D. Flynn, F. Dinmohammadi, M. Fisher, and M. Webster. 2019. Probabilistic model checking of robots deployed in extreme environments. arXiv:1812.04128v3 [cs.AI], 15 Feb 2019.

[50] X.-Y. Zhou and G. Z. Yang. 2019. Normalization in training u-net for 2d biomedical semantic segmentation. IEEE Robotics and Automation, 4, 2019, pp. 1792–1799.

[51] J. Lee, H. Davari, J. Singh, and V. Pandhare. 2018. Industrial artificial intelligence for industry 4.0-based manufacturing systems. Manufacturing Letters, 18, Oct. 2018, pp. 20–23.

[52] M. Prosser and J. D. Rebolledo. AI is kicking space exploration into hyper drive - here's how? Singularity hub, 07 Oct. 2018. Accessed on 21/10/2021.

[53] Y. Zhang, J. Xu, and L. Chen. Design of terrain recognition system based on laser distance sensor. Laser and Infrared, 46(3), 2016, pp. 265–270.

[54] Q. Zhang, X. Yang, and T. Liu. 2013. Design of a smart visual sensor based on fast template matching. Chinese Journal of Sensors and Actuators, 26(8), 2013, pp. 1039–1044.

[55] M. Wang, Y. Fan, and X. Wang. 2016. Design of infrared FPA detector simulator. Laser and Infrared, 46(12), 2016, pp. 1481–1485.

[56] X. Cui, Z. Wang, and C. Hou. 2015. Analysis and countermeasures to the problem of ultrasonic sensor receives the ultrasonic signal asymmetric. Chinese Journal of Sensors and Actuators, 28(1), 2015, pp. 1–85.

[57] H. Ouarda. Neural path planning for mobile robots. Int. J. of Systems Applications, Engineering & Development, 5(3), 2011, pp. 367– 376.

[58] H. He, T. M. McGinnity, Coleman S.A., and B. Gardiner. 2014. Linguistic decision making for robot route learning. IEEE Transaction on Neural Networks and Learning Systems, 25(1), 2014, pp. 203 – 215.

[59] R. Huang, H. Laing, and J. Chen. 2016. Lidar based dynamic obstacle detection, tracking and recognition method for driverless cars. Robot, 38(4), 2016, pp. 437–443.

[60] P. Wu, Y. Cao, Y. He, and D. Li. Computer Vision Systems. ICVS 2017. Lecture Notes in Computer Science, vol. 10528, chapter Vision-Based Robot Path Planning with Deep Learning. Springer, Cham, 2017.

[61] Z. Zhu, G. Xu, H. He, J. Jiang, and T. Wang. Recognition of speed signs in uncertain and dynamic environments. Journal of Physics: Conference Series, Application of computer network and information technology, 1187(4), Apr. 2019, 042066.

[62] J. C. Mohanta and A. Keshari. A knowledge based fuzzy probabilistic roadmap method for mobile robot navigation. Applied Soft Computing Journal, 79, 2019, pp. 391–409.

[63] H. He, T. Watson, C. Maple, A. Tiwari, J. Mehnen, Y. Jin, and B. Gabrys. The security challenges in the IoT enabled cyber-physical systems and opportunities for evolutionary computing & other computational intelligence. In proceedings of WCCI2016, Vancouver, Canada, 2016.

[64] A. Girard and C. Rommel. Industry 4.0: Secure by Design, 2017 VDC Research Group, Inc. Research Insights of Connected World. https://valbrio.com/wp-content/uploads/.

[65] H. Fang, A. Qi, and X. Wang. Fast authentication and progressive authorization in large-scale iot: How to leverage AI for security enhancement? ArXiv, abs/1907.12092, July 2019. IEEE Network, 34(3), June 2020, pp. 24-29, DoI: 10.1109/MNET.011.1900276.

[66] M. A. Salitin and A. H. Zolait, The role of User Entity Behavior Analytics to detect network attacks in real time, *the International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT),* Sakhier, Bahrain, 2018, pp. 1-5. DoI: 10.1109/3ICT.2018.8855782.

[67] H. Liu and B. Lang. Machine learning and deep learning methods for intrusion detection systems: A survey. Applied Sciences, 9(20), 2019, 4396. DoI: 10.3390/app9204396.

[68] A. Mason, Y. Zhao, H. He, R. Gompelman, and S. Mandava. Online anomaly detection of time series at scale. In IEEE International Conference on Cyber Situational Awareness, Data Analytics and Assessment (Cyber SA), Oxford, UK, 3-4 Jun. 2019.

[69] S. Wan, Y. Li, and K. Sun. Pathmarker: protecting web contents against inside crawlers. Cybersecurity, 2(1), 2019, pp. 1–17.

[70] D. Ucci, L. Aniello, and R. Baldoni. Survey of machine learning techniques for malware analysis. Computers & Security, 81, Mar 2019, pp. 123–147.

[71] A. Verma and V. Ranga. Machine learning based intrusion detection systems for IoT applications. Wireless Personal Communications, 4, 2020. DoI: 10.1007/s11277-019-06986-8.

[72] M. Bromiley. Empowering incident response via automation. SANSTM Institute, 22 Mar. 2019. https://www.sans.org/webcasts/.

[73] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In IEEE Symposium on Security and Privacy (SP), 2017, pp. 39–57.

[74] K. Ren, T. Zheng, Z. Qin and X. Liu, Adversarial Attacks and Defenses in Deep Learning, Engineering 6, 2020, pp. 346-360. DoI: 10.1016/j.eng.2019.12.012.

[75] Z. Gao, C. Cecati, and S. X. Ding. A survey of fault diagnosis and fault-tolerant techniques—part ii: Fault diagnosis with knowledge-based and hybrid/active approaches. IEEE Transactions on Industrial Electronics, 62, 2015, pp. 3768–3774.

[76] M. M. Hashimoto, H. Kawashima, and F. Oba. A multi-model-based fault detection and diagnosis of internal sensors for mobile robots. In Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003) (Cat. No.03CH37453), Las Vegas, NV, USA, 3 Dec. 2003.

[77] I. Eski, S. Erkaya, S. Savas, and S. Yildirim. Fault detection on robot manipulators using artificial neural networks. Robotics and Computer-Integrated Manufacturing, 27(1), Feb. 2011, pp. 115–123.

[78] C.N. Cho, J.T. Hong, and H.J. Kim. Neural network based adaptive actuator fault detection algorithm for robot manipulators. Journal of Intelligent & Robotic Systems, 95, 2019, pp. 137–147.

[79] Y. Ran and X. Zhou and P. Lin and Y. Wen and R. Deng, A Survey of Predictive Maintenance: Systems, Purposes and Approaches, ArXiv, 2019, abs/1912.07383.

[80] S. Xu, A Survey of Knowledge-Based Intelligent Fault Diagnosis Techniques Journal of Physics: Conference Series, 1187(3), 1 April 2019.

[81] L. A. Zadeh, "Toward Human Level Machine Intelligence - Is It Achievable? The Need for a Paradigm Shift," in *IEEE Computational Intelligence Magazine*, 3(3), pp. 11-22, August 2008, DoI: 10.1109/MCI.2008.926583.

[82] Q. Zhang and Q. Y. Yao. Dynamic Uncertain Causality Graph for Knowledge Representation and Reasoning: Utilization of Statistical Data and Domain Knowledge in Complex Cases. IEEE transactions on neural networks and learning systems 29(5), 2018, pp. 1637-1651.

[83] E. P. Nadeer, S. Mukhopadhyay and A. Patra, Hybrid System Model Based Fault Diagnosis of Automotive Engines. In: Sayed-Mouchaweh M. (eds) Fault Diagnosis of Hybrid Dynamic and Complex Systems. 2018. Springer, Cham.

[84] Y. Dai, M. Hinchey, M. Madhusoodan, J. L. Rash, and X. Zou. A prototype model for self-healing and self-reproduction in swarm robotics system. In 2nd IEEE Int. Symposium on Dependable, Autonomic and Secure Computing, Indianapolis, IN, USA, 29 Sept.-1 Oct. 2006.

[85] W. Xu. Towards human-centered AI: A perspective from human-computer interaction. INTERACTIONS, Jul-Aug. 2019.

[86] K.E. Vanderbilt, D. Liu & G. D. Heyman. The development of distrust. Child Dev. 82, 2011, pp. 1372 – 1380. DoI:10.1111/j.1467-8624.2011.01629.x.

[87] S. Vinanzi, M. Patacchiola, A. Chella, and A. Cangelosi. Would a robot trust you? Developmental robotics model of trust and theory of mind. Philosophical Trans. of the Royal Society B, 374(1771), 20180032, 2019.

[88] M. Patacchiola, & A. Cangelosi. A developmental cognitive architecture for trust and theory of mind in humanoid robots. IEEE Transactions on Cybernetics, 08 July 2020, pp. 1-13. DoI: 10.1109/TCYB.2020.3002892.

[89] C. Dio, F. Manzi, G. Peretti, A. Cangelosi, P.L. Harris, D. Massaro & A. Marchetti . Shall I trust you? From child–robot interaction to trusting relationships. Front. Psychol. 2020, DoI:10.3389/fpsyg.2020.00469.

[90] T. Young, D. Hazarika, S. Poria, and E. Cambria. Recent trends in deep learning based natural language processing. IEEE Computational Intelligence Magazine, 2018.

[91] A. Krieger. Industry solutions: Smart robot performs vision-assisted surgery, 15 May 2017. Accessed on 21/101/2021.

[92] J. Ruiz-del-Solar, P. Loncomilla & N. Soto, A Survey on Deep Learning Methods for Robot Vision, https://arxiv.org/pdf/1803.10862.pdf.

[93] M. Fritzsche, N. Elkmann and E. Schulenburg, Tactile sensing: A key technology for safe physical human robot interaction, 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Lausanne, Switzerland, 8-11 March 2011. DoI: 10.1145/1957656.1957700.

[94] A. Cirillo, P. Cirillo, G. D. Maria, C. Natale, and S. Pirozzi, A Distributed Tactile Sensor for Intuitive Human-Robot Interfacing, Journal of Sesnors, Vol. 2017, Article ID 1357061, DoI: 10.1155/2017/1357061.

[95] J. Gao and C. Zhou, A reasoning system about knowledge extraction in human-computer interaction, 2016 Chinese Control and Decision Conference (CCDC), Yinchuan, 2016, pp. 5450-5455, doi: 10.1109/CCDC.2016.7531971.

[96] S. N. Tran and A. S. d'Avila Garcez. Deep Logic Networks: Inserting and Extracting Knowledge from Deep Belief Networks. IEEE Trans Neural Networks and Learning Systems, 29(2), Feb. 2018, pp. 246-258. DoI: 10.1109/TNNLS.2016.2603784.

[97] M. Moore and S. Rugaber. Using knowledge representation to understand interactive systems. In Proc. Fifth Int. Workshop on Program Comprehension (IWPC'97), Dearborn, MI, USA, 1997, pp. 60-67.

[98] L. Atymtayeva. Automation of HCI engineering processes: System architecture and knowledge representation. Advanced Engineering Technology and Application, 4(2), 2015, pp. 41-46.

[99] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[100] L. El Hafi, S. Isobe, Y. Tabuchi, Y. Katsumata, H. Nakamura, T. Fukui, T. Matsuo, G.A. Garcia Ricardez, M. Yamamoto, A. Taniguchi, Y. Hagiwara, and T. Taniguchi, System for augmented human-robot interaction through mixed reality and robot training by non-experts in customer service environments, Advanced Robotics, 34(3-4), 2020, pp.157-172. DoI: 10.1080/01691864.2019.1694068

[101] H. He and J. Lawry, Linguistic Attribute Hierarchy and Its Optimisation for Classification Problems, Soft Computing, 18(10), Oct 2014, pp 1967-1984.

[102] H. He and J. Lawry, A Linguistic CMAC Equivalent to a Linguistic Decision Tree for Classification, *Int. Joint Conference on Neural Networks*, Atlanta, Georgia, USA, 14-19, Jun. 2009, pp. 1177 – 1183.

[103] S. Grossberg, A Path Toward Explainable AI and Autonomous Adaptive Intelligence: Deep Learning, Adaptive Resonance, and Models of

Perception, Emotion, and Action, Frontiers in Neurorobotics, 25 June 2020. DoI: 10.3389/fnbot.2020.00036.
[104] G. A. Carpenter, S. Grossberg, and J. H. Reynolds, Supervised Real-Time Learning and Classification of Nonstationary Data by a Self-Organizing Neural Network, Neural Networks, 4(5), 1991, pp. 565-588. DoI: 10.1016/0893-6080(91)90012-T.
[105] L. Fridman, Human-Centered Autonomous Vehicle Systems: Principles of Effective Shared Autonomy, arXiv:1810.01835v1 [cs.AI] 3 Oct 2018.
[106] T. Zhang, D. Tao, X. Qua, X. Zhang, R. Lin, and W. Zhang. The roles of initial trust and perceived risk in public's acceptance of automated vehicles. Transport Research Part C, 98, 2019, pp. 207–220.
[107] H. He, A. Ertan, R. N. Akram1, R. Hopcraft and H. Mansor, EAI Endorsed Transactions, invited book chapter Autonomous Vehicles - Cybersecurity and Privacy Challenges and Opportunities. IET, 2020. In press.
[108] K. Koscher, A. Czeskis, F. Roesner, S. Patel, T. Kohno, S. Checkoway, D. Mccoy, B. Kantor, D. Anderson, H. Shacham, and S. Savage. Experimental security analysis of a modern automobile. In IEEE Symposium on Security and Privacy, pages 447–462, Claremont Resort, Berkeley, CA. 2011.
[109] R. Abbas, S. Marsh and K. Milanovic, Ethics and System Design in a New Era of Human–Computer Interaction, IEEE Technology and Society Magazine, 38(4), Dec. 2019, pp. 32-33.
[110] Anand Rao, The real meaning of artificial intelligence, PwC, Recode, accessed on 19 Aug. 2020.
[111] A. Antonietti , D. Martina , C. Casellato , E. D'Angelo , and A. Pedrocchi, Control of a Humanoid NAO Robot by an Adaptive Bioinspired Cerebellar Module in 3D Motion Tasks, Computational Intelligence and Neuroscience Vol. 2019, Article ID 4862157. DoI: 10.1155/2019/4862157.
[112] S. Lukosch, Designing for Augmented Humans and Intelligence, *IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, Porto, Portugal, 2019, pp. 3-3, doi: 10.1109/CSCWD.2019.8791919.
[113] J. Lambert and E. Cone. How Robots Change the World: What Automation Really Means for Jobs and Productivity, Oxford Economics, Jun. 2019.
[114] G. Carriço, The EU and artificial intelligence: A human-centered perspective, European View, 17(1), 2018, pp. 29–36. DoI: 10.1177/1781685818764821
[115] AI for Good Global Summit, ITU in Geneva, Switzerland 7-9 June 2017. Access on 21/10/2021.
[116] The partnership on AI brings together diverse, global voices to realise the promise of Artificial Intelligence. Access on 21/10/2021.
[117] I. Asimov, "Runaround". *I, Robot* (The Isaac Asimov Collection ed.). New York City: Doubleday. 1950. p. 40. ISBN 978-0-385-42304-5.
[118] R. Murphy and D. D. Woods, "Beyond Asimov: The Three Laws of Responsible Robotics," in IEEE Intelligent Systems, 24 (4), pp. 14-20, July-Aug. 2009, DoI: 10.1109/MIS.2009.69.

Dr Hongmei He (SIEEE'16, FHEA) is currently an Associate Professor in Cybersecurity at De Montfort University. She obtained her Ph.D. degree in Computer Science from Loughborough University, UK in 2006. Her current research focuses on AI for the safety and security of Robotics & Autonomous System. She actively serves for IEEE UK & Ireland RAS Chapter as the chapter secretary and is the chair of the task force, "AI and Edge Computing for Trustworthy Robots and Autonomous Systems", in ADPRLTC of IEEE Computational Intelligence Society.

John Gray (MIEEE,FIET) is an Emeritus Professor of Robotics and Systems Engineering at the University of Manchester. In 1988 he established the UK's National Advanced Robotics Centre, served as its Research Director and as a member of the managing industrial consortium ARRL. Prof Gray has subsequently been involved in a range of European Commission and industrial funded robotics research projects. In 2000 he was invited by Maff/Defra to establish and chair the Food Manufacturing Engineering Group (FMEG), He has subsequently been involved in a large number of funded automation developments in this sector. He is currently an honorary editor of the Transactions on the Instruments of Measurement and Control and chair of the IEEE UK& Ireland RAS Chapter.

Angelo Cangelosi (SIEEE) is Professor of Machine Learning and Robotics at the University of Manchester (UK). He also is Turing Fellow at the Alan Turing Institute London, Visiting Professor at Hohai University and at Universita' Cattolica Milan, and Visiting Distinguished Fellow at AIST-AIRC Tokyo. His research interests are in developmental robotics, language grounding, human robot-interaction and trust, and robot companions for health and social care. His latest book "Cognitive Robotics" (MIT Press), coedited with Minoru Asada, will be published in 2021.

Qinggang Meng (SMIEEE) is currently a Professor in Robotics and AI with the Department of Computer Science, Loughborough University, UK. He is a fellow of the Higher Education Academy, UK. His research interests include biologically inspired learning algorithms and developmental robotics, service robotics, agricultural robotics, robot learning and adaptation, multi-UAV cooperation, human motion analysis and activity recognition, activity pattern detection, pattern recognition, artificial intelligence, computer vision, and embedded intelligence.

Prof T. Martin McGinnity (SMIEEE, FIET) received a Ph.D. degree from the University of Durham, UK in 1979. He currently holds a joint professorship at Nottingham Trent University (NTU) and Ulster University (UU), UK. Previously, he was Pro Vice Chancellor and Head of the College of Science and Technology, Dean of the School of Science and Technology at NTU, Head of the School of Computing and Intelligent Systems; Director of the Intelligent Systems Research Centre at UU. He is the author or co-author of 350+ research papers and leads the Computational Neuroscience and Cognitive Robotics research group at NTU. He is interested in industrial robotics, data analytics and medical systems.

Prof. Jörn Mehnen (MIEEE) research interests at the University of Strathclyde, UK, focus on Industry 4.0 and Digital Manufacturing. He is concerned with the conversion of deep academic insights into industrially highly applicable knowledge, skills, and technologies. His research covers Trustworthy IIoT, AI and Robotics, Additive Manufacturing, Cloud Manufacturing, Mixed Reality and Digital Twins for Manufacturing.