The challenges and opportunities of artificial intelligence for trustworthy robots and autonomous systems

The Challenges and Opportunities of Artificial Intelligence in Implementing Trustworthy Robotics and Autonomous Systems

H. He[1]* *Senior Member, IEEE,*, J. Gray† *Senior Member, IEEE,* Angelo Cangelosi †, *Senior Member, IEEE,* Q. Meng ‡ *Senior Member, IEEE,* T. M. McGinnity § *Senior Member, IEEE,* J. Mehnen ¶
*School of Aerospace, Transport and Manufacturing, Cranfield University, Cranfield, UK, MK43 0AL.
† Department of Electrical and Electronic Engineering, University of Manchester, Manchester, UK, M1 3BB.
‡Department of Computer Science, Loughborough University, Loughborough, UK, LE11 3TU.
§School of Science & Technology, Nottingham Trent University, Clifton, Nottingham NG11 8NF.
¶Design, Manufacturing and Engineering Management, University of Strathclyde, Glasgow, UK, G1 1XJ.

***Abstract*—Effective Robots and Autonomous Systems (RAS) must be trustworthy. Trust is essential in designing autonomous and semi-autonomous technologies, because "No trust, no use". RAS should provide high quality of services, with the four key properties that make it trust, i.e. they must be (i) robust for any health issues, (ii) safe for any matters in their surrounding environments, (iii) secure for any threats from cyber spaces, and (iv) trusted for human-machine interaction. We have thoroughly analysed the challenges in implementing the trustworthy RAS in respects of the four properties, and addressed the power of AI in improving the trustworthiness of RAS. While we put our eyes on the benefits that AI brings to human, we should realise the potential risks that could be caused by AI. The new concept of human-centred AI will be the core in implementing the trustworthy RAS. This review could provide a brief reference for the research on AI for trustworthy RAS.**

***Index Terms*—Artificial Intelligence, Trustworthiness of RAS, Cyber Security, Safety, Health, Human-Robot Interaction, Performance of RAS.**

I. Introduction

Artificial Intelligence (AI) and Machine Learning (ML) are increasingly used in robots and autonomous systems (RAS) that attempt to mimic the adaptive and smart problems solving capabilities of humans. Such systems promise a smarter and safer world —- where self-driving vehicles can reduce the number of road accidents, medical robots perform intricate surgeries, and "digital" pilots participate in crew flight-operations [1].

Now, RAS are becoming ubiquitous in different application domains, such as aerospace, transport, manufacturing, agriculture, social healthcare, and extreme environments, etc. Common important examples are robots, autonomous vehicles, unmanned aerial vehicles (UAV), autonomous trading systems, self-managing telecommunication networks, smart factories, and infrastructure, etc.

The Internet of Things (IoT) delivers new value by connecting People, Process and Data. Sensing and data analysis technologies in IoT are used to give robots a wider situational awareness that leads to better task execution. Hence, IoT technology could inspire wider applications of RAS. However, many RAS are released into the world without full prior analysis of potential inappropriate operations and may accomplish things which are not foreseen by their human designers or owners. As digital technologies are driving the autonomy of systems, there might be a risk that system autonomy could devalue human's work, thus giving rise to negative attitudes towards technology, or eventually leading to mass unemployment. Therefore, trustworthiness is particularly important on the way of implementing fully autonomous systems.

Humans tend to be always optimistic with regards to the potential of new techniques, and ignore or are initially unaware of the potential negative impact behind the advanced technologies. During the initial deployment of RAS, humans have tended to accept the untrusted products and services, but have gradually come to realise that Autonomous Systems Must Be Trustworthy. Many lessons have shown that trust directly influences operators' use of automation. For example, an autonomous vehicle killed a person in the street in Arizona, US in 2018, successful cyber-attacks have been executed to demonstrate how autonomous vehicles could potentially be hijacked, and the failure of an intelligent support aviation system on Boeing planes was responsible for the crash of two airplanes killing 189 and 157 people, respectively. Some intelligent robotic systems have catastrophically failed in situations where they were supposed to provide a high level of safety. Indeed, unforeseen and undesirable events can have significant negative impact on the acceptability of autonomous systems. This is not just a technical question as advances in autonomous systems, but also has ushered in a range of increasingly urgent and complex moral questions, and posed many ethical, societal and legal challenges.

[1]Hongmei He is the corresponding author, maryhhe@gmail.com

II. KEY FACTORS THAT AFFECT RAS' TRUSTWORTHINESS

Substantial strategic efforts on the trustworthiness of RAS is being made internationally, such as US National Institute of Standards and Technology (NIST) provides a trustworthiness framework of cyber physical systems, which covers cyber security, privacy, safety, reliability, and resilience. However, the functionality and the performance of RAS is essential for the acceptance of RAS, which represents the worthiness of RAS. A systematic approach is needed to co-create trust and worthiness of RAS, and the following aspects of trustworthiness of RAS need to be investigated thoroughly, thus promoting the implementation of fully trustworthy RAS.

Functionality/performance needs to well support the system autonomy from sensing, data collection and process, decision making, communication, human-machine interaction, action control and monitoring. Their logicality, performance and quality is the essential requirement of autonomous systems, and they represent the worthiness of a system. The functionality and performance of RAS needs to be guaranteed by the following properties of RAS, which could affect the trustiness of RAS, thus linking the acceptance of users to RAS.

Security is a critical challenge, as now everything goes to the Internet. Cyber-attacks could directly threaten the safety of an autonomous system. This has been demonstrated by the attack experiments on the SUV Jeep in 2015. Hence, RAS should be able to detect, defend and prevent any anomalies from cyberspace. The privacy of RAS is a branch of security, which focuses on the data protection, especially, personal information protection, and regulation compliance (e.g. GDPR). Security by Design and Privacy by Design is the requirements of industry 4.0.

Safety is a persistent requirement to all kinds of autonomous systems. Different application domains of autonomous systems may have different requirements in safety. For those systems (e.g. aircraft, vehicles, infrastructure), safety is a critical requirement; Safety is directly related to the reliability of RAS, which is an important factor to be concerned when human selects a kind of RAS. Safety should be co-designed with security, internal health and external interaction with human and environment.

Health, except the threats from cyber space and external environments, the reliability and the safety of RAS are facing the threats from potential process abnormalities and component faults. A fault is defined as an unpermitted deviation of at least one characteristic property or parameter of a system from the acceptable/usual/standard condition. The faults of RAS can be classified as actuator faults, sensor faults, and plant faults (or called component faults or parameter faults). Hence, it is paramount to detect and identify the diversity of potential abnormalities and faults as early as possible and implement fault tolerant operations for minimizing performance degradation and avoiding dangerous situations [13].

Human-Machine Interaction refers to the communications and interactions between a human and a machine via a user interface. In the NIST concept model of CPS (Fig. 2), the input of human can be fed into the decision loop. No matter which level a system's autonomy is at, human should be able to interrupt the system, even if the execution of autonomous systems does not need to have the intervention of human, and human interaction should be built upon tangible and attentive user interface principles. Through the intuitive and efficient visualization of sensed quantities, estimated statistics, and the automatic identification of trends over time, both users and the system could be better informed on their conditions and keep the consistency of the systems in the surrounding environments, thus increase their self-efficacy and trigger decision making. Especially, the last two principles of Norman's seven principles for HCI design [31], "Design for Error" and "When all else fails, standardize", should be applied to ensure the system to keep the bottom of safety and reliability.

Acceptance model of trustworthy RAS Zhang et al. [46] believed that security is not an important factor that affects the acceptance of CAVs, hence, the dependent relation from security to trust is depicted with a dashed arrow. Ertan et al. [10] argued that security could directly affect the reliability and the privacy, as highlighted by a Denial of Service (DoS) attack, which utilises a flood of arbitrary packets to a target system, and could cause a system malfunction, thus leading to a stuck or inoperable throttle, and raising potential safety concerns [23]. Therefore, cyber security is directly related to safety, but indirectly related to the trustiness. In fact, all the four properties in health maintenance, human-machine interaction, safety and cyber security are directly related to the reliability of RAS, which is directly related to the trustiness of RAS. Human-machine interaction and safety are directly related to the trustiness of RAS. Privacy depends on the protection of cyber security, but is directly related to the trustiness of RAS. The Health of RAS could affects the human-machine interaction. The reliability of RAS could affect the functionality and performance of RAS. An acceptance model is proposed in Fig. 1. Different application domains may have different weights of relations between different blocks in the model.

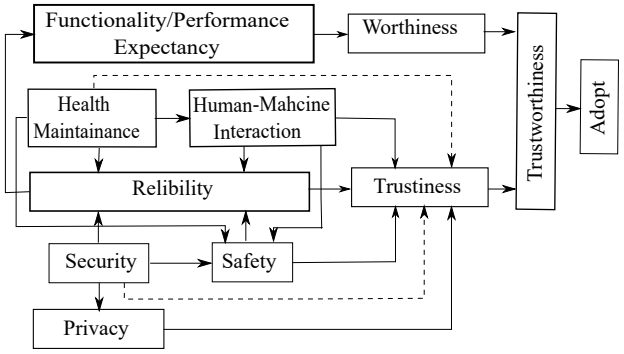

Fig. 1. Acceptance model for connected and autonomous vehicles

III. The challenges in implementing trustworthy RAS

A central function of an autonomous system is to enable it to use information that represents the state of the physical world and the cyber world, make decisions, and take actions, thus implementing the specific tasks with optimal performance and quality. The diversity, uncertainty and complexity of tasks as well as its cyber and physical environments always bring a big challenge to the optimisation of RAS performance. The most challenging goal is that the performance and quality of RAS for the specific tasks in the context of a specific application domain needs to align with the four properties of a system for its trustiness.

A. Challenges of RAS security

The IoT is where the Internet meets the physical world. There have been some serious implications on security as the attack threat moves from manipulating information to controlling actuation (i.e, moving from the digital to the physical world) [18]. Consequently, it drastically expands the attack surface from known threats and known devices in the upper layers of the IoT system stack, to additional security threats of RAS, communication protocols, and work processing of RAS. The attack surface of RAS is increased with more attack points that are vulnerable for cyber attackers, and through these attack points, attackers could intrude a system, by injecting data to, or extracting data from the system, thus compromising the security control of the system.

Katzenbeisser et al. [21] broadly divided the security of autonomous systems into security of the platforms that constitute it, including the security of hardware and software, and the security of communication that happens between these platforms. But they did not concern the threats from developmental platforms of RAS, which belong to the security issues on the supply chains of RAS [18].

Securing communication links is challengeable. There are different communication channels, such as wifi, GPS, Radio, bluetooth, etc. various attacks or crimes could intrude to RAS through these channels, for example, general Trojan-horse attacks on quantum-key-distribution systems, i.e., attacks on Alice or Bob's system via the quantum channel, peer-to-peer attacks on the same access point, MAC spoofing, wireless hijacking, denial of services, malicious eavesdropping, and Key Negotiation of Bluetooth (KNOB) Attacks, etc.

Securing the RAS software integrity is a critical challenge. Various attacks can break the integration of RAS software, thus producing different consequences, such as code modification, malfunctions, lose of the control, lose of customers' personal information, communication broken, high network traffic, etc. Among these, malfunctions and lose of control are critical challenges, as they could produce more severe consequences, and could directly break the safety of RAS and the customers who are using the RAS.

Securing hardware is a critical challenge as well. In an autonomous vehicle, many Embedded Computing Units (ECUs) could become an attack point. Securing these ECUs is a critical challenge. As an ECU has limited computing resources, it is difficult to have a comprehensive and effective security solution on it with a real-time performance. Side channel attacks (SCAs) are typical attacks on embedded systems, aiming at collecting leakage data during processing and through statistical computations extract sensitive information. For example, a Rowhammer fault injection attack can be mounted even remotely to gain full access to a device DRAM (Dynamic Random Access Memory); Cache side-channel attacks retrieve secret information by monitoring the cache of the system. Moreover, an attacker could intrude the sensing system of RAS, thus producing adversarial readings to a data collection system, which may cause a wrong decision, and further lead to a wrong action.

B. Challenges of RAS safety

NIST [15] proposed a concept model of cyber physical systems (CPS), as shown in Fig. 2. All RAS belong to CPS. Hence, the CPS model is applicable to RAS. The safety

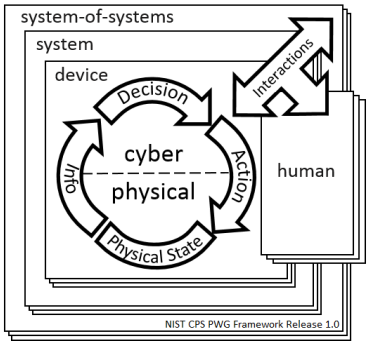

Fig. 2. The NIST model of cyber-physical systems [15]

of RAS covers two aspects: the safe threat to RAS from surrounding environments and the safe threat to surrounding environments from RAS. These two aspects of safety are closely linked with the sensors and actuators in RAS. In the loop of the CPS model, the information from the sensing system of RAS is important, as it represents the physical state of the surrounding environment, and is fed into the decision system. In a non-fully autonomous system, the information may also come from an operator. The decision phase receives the information from the sensing system or operators, and generates some plans according to an abstract representation of the system and its environment. For example, in a robot's navigation system, the decision output is the optimal path on which the robot can avoid any obstacles and reach to the target. This requires the robot to be able the detect obstacles on the way towards the target. The decision functions could be planning,learning or goal reasoning, depending the specific application domain. Hence, to ensure the safe process or work of RAS, RAS should be able to overcome the following challenges:

(1) Various working environments, which require RAS to be able to rightly sense the surrounding environments, deal with any anomalies in the environments and make decision that aligns with the goal of the RAS in real-time. For example, autonomous vehicles could face various anomaly situations, such as pedestrians crossing the road suddenly, an accident in front of the vehicle, obstacles in front of the vehicle, and road direction changes, etc.

(2) Various tasks, which require RAS (e.g. robots) to be able to synchronise the actions of human and other robots in a collaborative team, and avoid any disorders, which may lead the robot to hit other members (human or robots) in the team.

(3) Diversity and uncertainty of potential failures, which requires RAS to be effective and efficient for detecting any failures, early warning and fast responding to the detected failures of the system. In an extreme case, RAS should allow human's intervention at any execution point.

C. Challenges of RAS health

A fault is occurring when there is a difference between the realizable function and the required function. Kawabata et al. [22] categorised three types of faults: fatigue (deterioration), noise (sudden fault) and initial failure. Predictive maintenance may be more important in the RAS domain than in other domains, as an autonomous system without human intervention could be more severely damaged by a fault in the system. Diagnostics is concerned with current state of any subsystem whereas prognostic is related to the future state of subsystem [34]. Hence, for a fully autonomous system, prognostic may be more important, and it is challengeable to get high accurate prognostic, as it depends on the usage of the system, the experience of operators as well as the working environments, and hence, these factors might decide the fatigue level and speed of system components. A robot or an autonomous system is a complex system. For example, a vehicle has a very complex mechatronic structure consisting of subsystems, such as gearbox, engine, brakes, fuel, ignition, exhaust, and cooling. Normally any subsystem comprises electromechanical processes, actuators, and sensors. The sensors and actuators are associated and controlled with an ECU, which manages and screens the procedure. However, to save the cost, manufacturers often install limited sensors to monitor the system. Hence, the limit amount of data from sensors may limit the performance and the coverage of diagnosis. Also, online diagnostics and prognostics should align with the requirement of real-time performance, which is a critical challenge.

Generally, a self-diagnosis system for an autonomous system can consist of three processes: internal condition sensing, diagnosis and coping with the faulty condition to increase the fault tolerance, plus fast responses for sudden faults. Faults from sensors could lead to a wrong decision in the system; faults from actuators may cause a wrong behviour of the system; and faults from electronic components could cause the malfunction and disorder of the system. The complexity, diversity and uncertainty of faults brings many challenges. Especially for complicated faults, it is challengeable to find the fault origin, which brings the challenge to implement fast and right responses.

D. Challenges of RAS interaction with human

Although we expect to implement fully autonomous systems, we still need to design the systems to allow human to interact with the systems, in case any emergent situations. The researchers in [36] realised that humans may want to interrupt a robot on its autonomous execution, but it seems very difficult without the pre-design of interruptions. The human-robot interaction for allowing a human to interrupt a robot is complex, as many situational features and constraints need to be considered, including task priorities, operations, interruption frequency, and timings.

There could be many applications, in which, human and robots work in a collaborative team. Especially, for those critical problem domains, such as defence, healthcare, and industry, where the consequences of mistakes, errors or failure to perform are dire, RAS could be applied to replace human to deal with dangerous, difficult or complex cases. However, as our physical environments are dynamic, non-deterministic, and partially unknown, implementing trusted Human-Robot Interaction in such a complex physical world is a challenge.

Also, the communication with humans requires socially acceptable responses and common-sense knowledge to handle a broad variety of situations with complex semantics to interpret and understand. For example, a social robot needs to express, understand, and induce emotions as part of the interaction process. The diversity, complexity, and uncertainty of human status brings a big challenge, especially human emotion expression is complicated and uncertain, and even a kind of emotion could be expressed differently by different subjects.

Most importantly, the critical challenge lies in the inability to exhaustively test such a complex human-machine interaction system, especially the response, learning and adaptation of an autonomous system to unforeseen circumstances in a dynamic and changing world.

IV. The Opportunities of AI Technology

AI techniques have been applied in widespread areas, such as extreme environments, social-health care, manufacturing and military, etc. and the power of AI has been demonstrated in various applications for different purposes. For example, recently, Zhao et al. [48] proposed a probabilistic model to verify the safety and reliability of unmanned underwater vehicles in extreme environments; Zhou and Yang [49] investigated different types of nomalisation in training Deep Convolutional Neural Networks for 2D biomedical semantic segmentation; Lee et al. [25] proposed an industrial AI ecosystem; It was reported that AI is also being used for trajectory and payload optimization, which are important preliminary steps to NASA's next rover mission to Mars, the Mars 2020 Rover [33].

To improve the trustworthiness of RAS, the performance and the quality of services provided by RAS should incorporate with the trust properties, that is, RAS should be: robust for any health issues, safe for any matters in their surrounding environments, secure for any threats from cyber spaces, trusted for human-machine interaction.

A. AI for the security of RAS

Automation is the only way to level the playing field, reduce the volume of threats, and enable faster prevention. Naturally, AI is the key driver for the automation of RAS' cyber security. Unlike other problem domains, the design of intelligent solutions for cyber security has to be resilient in the face of determined and sophisticated attackers, who may target any kinds of RAS, which are usually connected to the Internet for improving their computing capacity, storage capacity, accessibility, usability and flexibility, etc. Cyber Intelligence is expected to be able to secure the benefits to all from the cyber-connected world. A mechanism is needed to allow the security components to be seamlessly integrated into the architecture of RAS, which should enable security to be adaptive, self-learning and autonomous, and thus to implement "Security by Design", demanded by Industry 4.0.

AI techniques have been applied in the tasks of security protection and attack detection. For access control, AI techniques have been applied for solving many authentication problems, such as bio-metrics identification (e.g. palm, iris, fingerprint and face), signature verification, keystroke pattern recognition, etc. For example, Fang et al. [12] developed new AI enabled security provisioning approaches to achieve fast authentication and progressive authorization.

Attack or intrusion detection is important for the cyber security of RAS. Machine Learning techniques have played important roles in data-driven cyber security, as they bring two significant gains to threat Intelligence: first, machines can deal with huge amount of data and their complex relations, which it is impossible to do with manpower alone; second, machines can implement the automation of cyber security, which it is not possible to implement by human. There has been much research in this area, such as intrusion detection [26], anomaly detection [28], crawler detection [38], malware analysis [35], and human behaviour monitoring [16], etc.

Improving detection accuracy, reducing the false alarm rate and detecting unknown attacks are everlasting goals for machine learning-based intrusion detection systems. Due to the development of attack techniques, adaptive Intrusion Detection Systems (IDS) are demanded to detect known and new attacks. IDS could passively operate to prevent impact on RAS. Most of IDS technologies operate at network level. For IoRT enabled systems, the capacity of intrusion detection in edge devices is needed. However, the computing constraints and real-time performance requirements of edge devices may limit the capacity of edge IDS. Therefore, the online intrusion detection of RAS is a critical challenge. Verma and Ranga [37] investigated machine learning classification algorithms for securing IoT against DoS attacks, and they used a Raspberry Pi system to evaluate the response time of classifiers on the IoT specific hardware.

With the growing number of cyber-attacks and increasingly complex IT environments, an intelligent incident response plan is more than just a set of instructions. The best solution of a fast response is to empower an incident response via automation [3], which could be implemented using social-technical model based on the effective and efficient threat intelligence, alert enrichment and a priority order of actions on RAS.

However, as discussed in [4], [5], machine learning (ML) models, as the driver of threat intelligence could be vulnerable to adversarial examples. When an ML model is trained with adversarial examples, it could over-fit to the adversarial examples, thus leading to wrong results at test. Therefore, one of challenges of deploying AI-based techniques to security domains is to solve the over-fitting problem. In real world applications of cyber security, it is difficult to check if the collected data is wrong or not. Namely, ML techniques themselves do represent a grand solution for the automation of cyber security, but requiring correct (or trustworthy) features or data to be available. In addition, an AI model, as a program in a system, could be hacked, like other programs in the system. This could produce unexpected consequences. Therefore, the robustness and security of data collection systems and ML models need to be investigated in the system design.

ML models can be evaluated with different performance indicators in terms of the tasks in the applications of RAS. Effectiveness (e.g. accuracy, F-measure, ROC curves, etc.) and efficiency (e.g. real-time) are highly required, while computational resource is restricted for the on-board countermeasures to secure RAS.

B. AI for the Safety of RAS

To insure the safety of RAS, a pervasive monitoring system is necessary. NASA updated conventional monitoring system architecture by including user inputs in the monitoring loop, as shown in Fig. 3. Many of RAS are critical safe systems

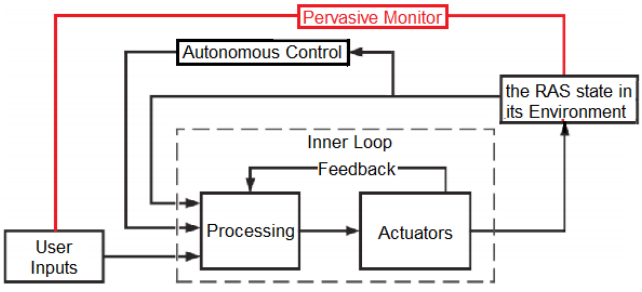

Fig. 3. Pervasive Monitoring Architecture,derived from [1]

(e.g. airplanes, autonomous vehicles), equipped with a Safety Instrumented System (SIS) for specific control functions to fail-safe or maintain safe operations of a process when unacceptable or dangerous conditions occur. SIS for different types

of autonomous systems could be implemented in different manners, but they are usually composed of sensors, logic solvers, actuators and other control equipment. One of key tasks in SIS is the detection of anomalies through sensing their surround environments. Advanced sensor technology greatly improves RAS' perception. The most frequently used sensors include laser sensors [17], [44], [47], visual sensors [45], radar, GPS, infrared sensors [40] and ultrasonic sensors [39], etc. A sensing system should function for data collection to enable system perception.

In all autonomous systems, aircraft could be required to have the highest requirements for safety. In terms of NASA's report [1], system monitoring plays an important role for the safety of RAS. For example, an aircraft could have many monitoring subsystems, such as instrument monitoring, system monitoring, and environment monitoring. Navigation is an important property of RAS, which infers the running status, adjusts the settings for flight appropriately, based on the information from all monitoring systems.

There has been much research on AI techniques, especially machine learning techniques, for environment monitoring and RAS navigation. For example, neural networks have been developed for a robot's path planning [20], [32]; a linguistic decision tree was developed for the classic robot routing learning problem [17]; a support vector machine based on the space-time feature vector was developed to recognize dynamic obstacles [19]; a Deep Convolutional Neural Network (DCNN) was developed for robot's navigation [41]; To improve the recognition rate of the speed signs for autonomous vehicles in dynamic environments, a Spatial Pyramid Pooling based DCNN was developed to recognise speed signs in dynamic environments based on the node in images extracted with the method of the salient target detection on the background-absorbing Markov chain for autonomous vehicles [50]; a knowledge-based fuzzy control system was developed for target search behavior and path planning of mobile robots [29]. However, online machine learning training to adapt the dynamic environment is still an open question. Improving the perception (e.g. obstacle detection), positioning accuracy, decision accuracy, and resilience to the surrounding environment are everlasting goals for RAS.

C. AI for the health of RAS

To improve the reliability of a system, fault diagnosis is usually employed to monitor, locate, and identify the faults. The analytical redundancy techniques have become the main stream of the fault diagnosis research since 1980s [14]. Fig. 4 illustrates the framework of an analytical fault diagnosis model, where, f_a is an actuator fault, f_c is a process/component fault, and f_s is a sensor fault, the input u and the output y are used to construct a fault diagnosis algorithm, which is employed to check the consistency of the feature information of the real-time process carried by the input and output data against the pre-knowledge on a healthy system, and a diagnostic decision is then made by using diagnostic logic. Now, with the development of AI techniques, the diagnosis algorithm can be implemented using a machine learning model, which can be trained by a set of historical data from the status information of sensors, actuators and components, including the fault information, f_a, f_c and f_s.

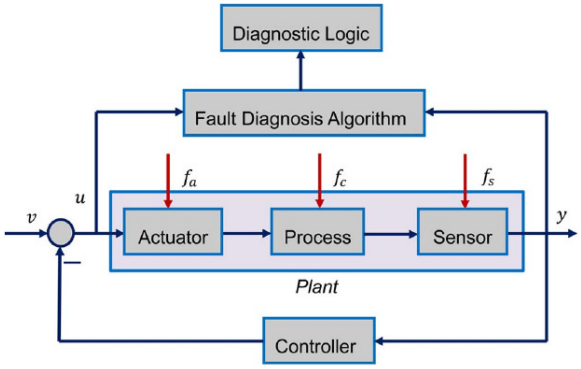

Fig. 4. Analytical Fault Diagnosis [14]

Usually, fault diagnosis includes three tasks, such as fault detection, fault isolation, and fault identification. Fault detection is the most basic task of the fault diagnosis, which is used to check whether there is a malfunction or a fault in the system and determine the time when the fault occurs, fault isolation is to determine the location of the faulty component, and fault identification is to determine the type, shape, and size of a fault. Fault diagnosis can be categorised to four types: model-based fault diagnosis, signal-based fault diagnosis, knowledge-based and hybrid methods. The idea of model-based diagnosis is to create a model to map the relationship between inputs and outputs (Eq. (1)), which can be represented by a machine learning model.

$$\hat{y} = \mathscr{L}(f(u)). \tag{1}$$

The system will check the output y is consistent with the model output $\hat{y}$, if no consistence, then the system is faulty. For example, Hashimoto et al. [27] developed an approach to the detection and diagnosis of the hard/noise failure based on the variable structure interacting multiple-model estimator, where, changes of the failure modes are modeled as switching from one mode to another in a probabilistic manner.

Signal-based methods utilize measured signals rather than explicit input–output models for fault diagnosis. It can be further divided to three types: time-domain, frequency-domain and time-frequency domain signal based fault diagnosis [13]. They are often used in a monitoring system. The faults in the process are reflected in the measured signals, whose features are extracted, and a diagnostic decision is then made based on the symptom analysis and prior knowledge on the symptoms of the healthy systems, which can be fed into an AI model to implement automatic diagnosis.

Recently, machine learning techniques have been applied for data-driven fault detection (as a decision maker) or diagnosis (as a classifier). The data can be the directly measured signals

or features extracted from signals or raw data from sensors. For example, Artificial Neural Networks have been developed for predicting the fault in terms of the vibration of a robot's joints [11], and estimating the fault torque for the adaptive actuator of a robot for robot manipulators [6]. A hybrid approach of combining knowledge-based model and machine learning model might benefit for the improvement of the precision of fault diagnosis for robot systems. Similar to the AI for data-driven IDS or anomaly detection, the three challenges in diagnosis accuracy, real-time performance, and data availability are applicable for data-driven fault diagnoses. Fault allocation is a challenge for a complex system, comprised of multiple components. Obviously, AI techniques in optimisation could be applicable for the fault allocation problem.

As a large-scale distributed system, a swarm robotic system represents a group of robots (e.g. drones) working for a mission with swarm intelligence for extreme and hazard environments or entertainments. In such a large swarming system without the intervention of human, autonomous self-diagnosis, self-healing and self-reproduction is necessary. Dai et al. [8] presented a self-healing and self-reproduction mechanism based on the virtual neurons with consequence-oriented prescription.

D. AI for trusted human-machine interaction

Human-machine interaction is a challenge for artificial intelligence. This field lays at the crossroad of several domains of AI and requires to tackle them in a holistic manner: modelling humans and human cognition; acquiring, representing, manipulating in a tractable way abstract knowledge at the human level; reasoning on this knowledge to make decisions; and eventually instantiating those decisions into physical actions both legible to and in coordination with humans. HCI researchers strive to leverage advanced technologies from AI and quantum computing, etc. into delightful, easy-to-use human-machine interaction systems that aligning with our lives. *Modelling human cognition* Natural Language Processing (NLP) is important technique for improving the cognition of human-machine interaction. For decades, machine learning approaches targeting NLP problems have been based on shallow models (e.g., Support Vector Machine (SVM) and logistic regression) trained with very high dimensional and sparse features. In the last few years, neural networks based on dense vector representations have been producing superior results on various NLP tasks [43]. NLP mainly includes the 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, for example, rule-based machine translation, and of analyzing different aspects of the language. NLG involves text planning, sentence planning and text realization.

Robot Vision enables more natural interaction with humans. By adding visual understanding capabilities to a robot, it can perceive human action and can naturally interact with humans through these non-verbal behaviours, like body gestures, facial expressions, and body poses. This requires the robot to have the capability of understanding these non-verbal behaviours, for which AI has played an important role. For example, a smart robot could be used to assist physicians in performing surgery, using near IR and 3D cameras [24], and a robot can recognise human's behaviour and emotion with AI and vision techniques.

Knowledge extraction and sharing between human and machines is a dynamic process of human-computer interaction, the input of a decision problem and the output of the solution will be converted into the knowledge that can be extracted by a machine. Collobert et al. [7] demonstrated that a simple deep learning framework outperforms most state-of-the-art approaches in several NLP tasks such as named-entity recognition, semantic role labeling, and Part-Of-Speech tagging.

Knowledge representation can be used to model the abstractions, and it has the advantage of providing support for transformation to the user interface environment [30]. Also, adaptive knowledge representation is important for the automation of HCI engineering processes [2]. Devlin et al. [9] proposed a new language representation model, bidirectional encoder representations from transformers, to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers.

V. Conclusions

We have identified the key factors that could significantly affect the trustworthiness of RAS, such as cyber security, safety, health, interaction of RAS with human. The performance/functionality of RAS represents the worthiness of RAS, subject to the four properties. We analysed the challenges of these properties related to the trustworthiness, and reviewed the power of AI techniques in the implementation of RAS trustworthiness. On the way towards the fully autonomous systems, human has realised that we must overcome many ethical, societal and legal challenges. AI has played significant roles in the development of trustworthy RAS. However, AI has both potential benefits and risks. Machine learning-based AI systems trained with incomplete or distorted data (i.e. their "worldview") can lead to biased "thinking," which may in turn magnify prejudice and inequality, spread rumors and fake news, and even cause physical harm. Hence, a new concept of human-centered AI (HAI) was raised in the AI community. It emphasizes that the next frontier of AI is not just technological but also humanistic and ethical with the three objectives [42]: (1) to technically reflect the depth characterized by human intelligence; (2) to improve human capabilities rather than replace them; and (3) to focus on AI's impact on humans.

References

[1] E. E. Alves, D. Bhatt, B. Hall, K. Driscoll, A. Murugesan, and J. Rushby. Considerations in assuring safety of increasingly autonomous systems. Technical Report NASA/CR–2018-220080, NASA, July 2018. accessed on 25 Jan 2020.

[2] L. Atymtayeva. Automation of hci engineering processes: System architecture and knowledge representation. *Engineering*, 2015.

[3] M. Bromiley. Empowering incident response via automation. SANS^{TM} Institute, 22 Mar. 2019. access on 25 Jan 2020.

[4] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou. Hidden voice commands. In *The 25th USENIX Security Symposium (USENIX Security 16)*, pages 513–530, 2016.

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

[6] 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*, 95:137–147, 2019.

[7] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, page 2493–2537, Aug. 2011.

[8] 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 International Symposium on Dependable, Autonomic and Secure Computing*, Indianapolis, IN, USA, 29 Sept.-1 Oct. 2006. IEEE.

[9] 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.

[10] A. Ertan, H. Mansor, H.* He, R. N. Akram1, and R. Hopcraft. *EAI Endorsed Transactions*, invited book chapter Autonomous Vehicles - Cybersecurity and Privacy Challenges and Opportunities. IET, 2019. in review.

[11] 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):115–123, Feb. 2011.

[12] 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.

[13] 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):3757–3767, Jun. 2015.

[14] 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*, 62:3768–3774, 2015.

[15] E. R. Griffor, C. Greer, D. A. Wollman, and M. J. Burns. Framework for cyber-physical systems: Volume 1, overview. accessed on 12 Mar. 2020.

[16] R. Haidar and S. Elbassuoni. Website navigation behavior analysis for bot detection. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Tokyo, Japan, 19-21 Oct. 2017. IEEE.

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

[18] 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 *WCCI2016*, Vancouver, Canada, 2016.

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

[20] I.K. Jung, K.-B. Hong, S.-K. Hong, and S. C. Hong. Path planning of mobile robot using neural network. In *IEEE International Symposium on Industrial Electronics(ISIE'99),*, volume 3, pages 979–983, Bled, Slovenia, 12-16 July 1999.

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

[22] K. Kawabata, S. Okina, T. Fujii, and H. Asama. 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*, 17:925–950, 2003.

[23] 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.

[24] A. Krieger. Industry solutions: Smart robot performs vision-assisted surgery. accessed on 12 Mar. 2020.

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

[26] H. Liu and B. Lang. Machine learning and deep learning methods for intrusion detection systems: A survey. *Applied Sciences*, 9(4396):1–28, 2019.

[27] 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. IEEE.

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

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

[30] M. Moore and S. Rugaber. Using knowledge representation to understand interactive systems. In *Proceedings Fifth International Workshop on Program Comprehension (IWPC'97)*, pages 60–67, Dearborn, MI, USA, 1997.

[31] D. A. Norman. *The Design of Everyday Things*. Doubleday, New York, 1988.

[32] H. Ouarda. Neural path planning for mobile robots. *Internaitonal Journal of Systems Applications, Engineering & Development*, 5(3):367–376, 2011.

[33] M. Prosser and J. D. Rebolledo. Ai is kicking space exploration into hyperdrive—here's how? Singularityhub, Oct. 2018. accessed on 12 Mar. 2020.

[34] 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*, 18:179–194, 2017.

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

[36] M. Veloso. A few issues on human-robot interaction for multiple persistent service mobile robots. In *AAAI 2014 FALL SYMPOSIUM SERIES*, Arlington, Virginia, USA, 2014.

[37] A. Verma and V. Ranga. Machine learning based intrusion detection systems for iot applications. *Wireless Personal Communications*, pages 1–24, Nov. 2019.

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

[39] Cui X. Wang, Z. and C. Hou. Analysis and countermeasures to the problem of ultrasonic sensor receives the ultrasonic signal asymmetric. *Chinese Journal of Sensors and Ac-tuators*, 28(1):81–85, 2015.

[40] M. Wang, Y. Fan, and X. et al. Wang. Design of infrared fpa detector simulator. *Laser and Infrared*, 46(12):1481–1485, 2016.

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

[42] W. Xu. Toward human-centered ai: A perspective from human-computer interaction. *INTERACTIONS*, Jul-Aug. 2019.

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

[44] D. Zhang, W. Li, and H. et al. Wu. Mobile robot adaptive navigation in dynamic scenarios based on learning mechanism. *Information and Control*, 45(5):521–529, 2016.

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

[46] 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*, 98:207–220, 2019.

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

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

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

[50] 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):042066, apr 2019.