ON SPECIFYING FOR TRUSTWORTHINESS

Dhaminda B. Abeywickrama
Department of Aerospace Engineering
University of Bristol, UK
dhaminda.abeywickrama@bristol.ac.uk

Amel Bennaceur
Department of Computing
Open University, UK
amel.bennaceur@open.ac.uk

Greg Chance
Department of Computer Science
University of Bristol, UK
greg.chance@bristol.ac.uk

Yiannis Demiris
Electrical and Electronic Engineering
Imperial College London, UK
y.demiris@imperial.ac.uk

Anastasia Kordoni
Department of Psychology
Lancaster University, UK
a.kordoni@lancaster.ac.uk

Mark Levine
Department of Psychology
Lancaster University, UK
mark.levine@lancaster.ac.uk

Luke Moffat
Department of Sociology
Lancaster University, UK
l.moffat1@lancaster.ac.uk

Luc Moreau
Department of Informatics
King's College London, UK
luc.moreau@kcl.ac.uk

Mohammad Reza Mousavi
Department of Informatics
King's College London, UK
mohammad.mousavi@kcl.ac.uk

Bashar Nuseibeh
Department of Computing
Open University, UK
b.nuseibeh@open.ac.uk

Subramanian Ramamoorthy
School of Informatics
University of Edinburgh, UK
s.ramamoorthy@ed.ac.uk

Jan Oliver Ringert
Department of Computer Science
Bauhaus University Weimar, Germany
jan.ringert@uni-weimar.de

James Wilson
Department of Engineering Mathematics
University of Bristol, UK
j.wilson@bristol.ac.uk

Shane Windsor
Department of Aerospace Engineering
University of Bristol, UK
shane.windsor@bristol.ac.uk

Kerstin Eder
Department of Computer Science
University of Bristol, UK
kerstin.eder@bristol.ac.uk

ABSTRACT

As autonomous systems are becoming part of our daily lives, ensuring their trustworthiness is crucial. There are a number of techniques for demonstrating trustworthiness. Common to all these techniques is the need to articulate *specifications*. In this paper, we take a broad view of specification, concentrating on top-level requirements including but not limited to functionality, safety, security and other non-functional properties. The main contribution of this article is a set of high-level intellectual challenges for the autonomous systems community related to specifying for trustworthiness. We also describe unique specification challenges concerning a number of application domains for autonomous systems.

Keywords Autonomous systems · Trust · Specification

1 Introduction

Autonomous systems (ASs) are systems that involve software applications, machines and people, which are capable of taking actions with no or little human supervision [1]. Soon, ASs will no longer be confined to safety-controlled industrial settings. Instead, they will increasingly become part of our daily lives having matured across different domains like driverless cars, healthcare robotics and uncrewed aerial vehicles.

Trust may vary, as it can be gained and lost over time. Different research disciplines define trust in different ways. This article focuses on the notion of trust that concerns the relationship between humans and ASs. ASs are considered *trustworthy* when the design, engineering, and operation of these systems generate positive outcomes and mitigate outcomes which can be harmful [2]. Trustworthiness of ASs can be dependent on many factors such as: (i) explainability, accountability and understandability to different users; (ii) robustness of ASs in dynamic and uncertain environments; (iii) assurance of their design and operation through verification and validation (V&V) activities; (iv) confidence in their ability to adapt their functionality as required; (v) security against attacks on the systems, users, and deployed environment; (vi) governance and regulation of their design and operation; and (vii) consideration of ethics and human values in their deployment and use [2].

There are various techniques for demonstrating trustworthiness of ASs, such as synthesis, formal verification at design time, runtime verification or monitoring, and test-based methods. But, common to all these techniques is the need to formulate *specifications*. According to the ISO standard for systems and software engineering vocabulary [3], a *specification* is a detailed formulation that provides a definitive description of a system for the purpose of developing or validating the system.

Writing specifications that inspire trust is challenging [4]. A human may trust an AS to perform its actions, if it demonstrably acts in an effective and safe manner. Thus, the AS not only needs to be safe and effective, but also needs to be perceived as such by humans. In the same manner, in human-robot interaction, it is equally important to ensure that the AS can trust humans. To address this, specifications will need to go beyond typical functionality and safety aspects. Thus, in this context, we take a broad view of specification, concentrating on top-level requirements including but not limited to functionality, safety, security and other non-functional properties that contribute to trustworthiness of ASs. Also, we intentionally leave the discussion on the formalisation of these specifications for the future, when understanding of what is required to specify for trustworthiness is more mature.

In this paper, we explore what it means to specify for trustworthiness, considering a number of ASs domains, each with its unique specification challenges. We then present a set of intellectual challenges for the ASs community comprising both academia and industry, specifically focusing on specification for trustworthy ASs.

2 Autonomous Systems Domains and their Specification Challenges

In this section, we discuss the specification challenges of a number of different ASs domains (see Fig. 1).

- **Automated Driving**

Automated driving (self-driving) refers to a class of ASs that vary in the extent to which they make decisions independently (SAE J3016 standard taxonomy). The higher levels of autonomy, levels 3-5, refer to functionality ranging from traffic jam chauffeur to completely hands-free driving in all conditions. Despite an explosion of activity in this domain in recent years, the majority of systems being considered for deployments depend on careful delineation of the operation design domain (ODD) to make the specification of appropriate behaviour tractable. Even so, the specification problem remains difficult for a number of reasons. Firstly, traffic regulations are written in natural language, ready for human interpretation. Although the highway code rules are intended for legal enforcement, they are not specifications that are suitable for machines. There typically are many exceptions, context-dependent conflicting rules and guidance of an 'open nature', all of these require interpretation in context. Driving rules can often be vague or even conflicting and may need a base of knowledge to interpret the rule given a specific context. The UK Highway Code Rule 163 states that after you have started an overtaking manoeuvre you should "move back to the left as soon as you can but do not cut in" [5]. A more explicit specification of driving conduct (e.g. Rule 163) to something more machine interpretable that captures the appropriate behaviour presents a challenge to this research area. Secondly, driving in urban environments is an intrinsically interactive activity, involving several actors whose internal states may be opaque to the automated vehicle. As an example, the UK Highway Code asks drivers to not "pull out into traffic so as to cause another driver to slow down". Without further constraint on what the other drivers could possibly do, specifying appropriate behaviour becomes difficult, and any assumptions made in that process would call into question the safety of the overall system when those assumptions are violated. Thus, two key challenges in the area of automated driving

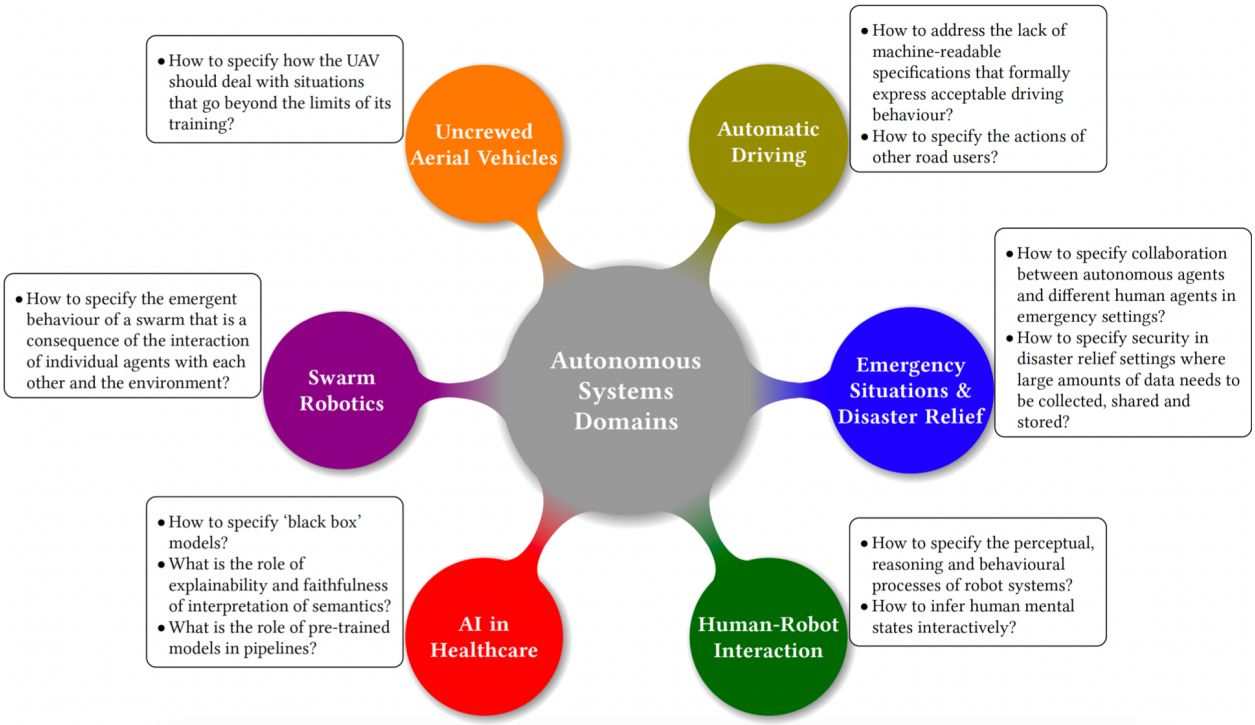

Figure 1: ASs domains and their specification challenges.

are the lack of machine-readable specifications that formally express acceptable driving behaviour and the need to specify the actions of other road users (see also Fig. 1).

- **Emergency Situations and Disaster Relief**

Emergency situations evolve dynamically and can differ in terms of the type of incident, its magnitude, additional hazards, the number and location of injured people. They are also characterised by urgency; they require a response in the shortest time frame possible and call for a coordinated response of emergency services and supporting organisations, which are increasingly making use of ASs. This means that successful resolutions depend not only on effective collaboration between humans [6], but also between humans and ASs. Thus, there is a need to specify both functional requirements and SLEEC requirements (the Social, Legal, Ethical, Empathic, Cultural rules and norms that govern an emergency scenario). This suggests a shift from a static design challenge towards the need to specify for adaptation to the diversity of emergency actors and complexity of emergency contexts. In addition, to enhance collaboration between autonomous agents and different human agents in emergencies, specifying human behaviour remains one of the main challenges in emergency settings.

There are also challenges for specifying security in the context of disaster relief. A large part of this comes from the vast amounts of data that needs to be collected, shared and stored between different agencies and individuals during an emergency scenario. Securing a collaborative information management system is divided between technical forms of security, such as firewalling and encryption, and social forms of security, such as trust. In order to properly provide security to a system, both aspects must be addressed in relation to each other within a specification.

- **Human-Robot Interaction**

Interactive robot systems aim to complete their tasks while explicitly considering states, goals and intentions of the human agents they collaborate with, and aiming to calibrate the trust humans have for them to an appropriate level. This form of human-in-the-loop real-time interaction is required in several application domains including assistive robotics for activities of daily living [7], healthcare robotics, shared control of smart mobility devices [8], and collaborative manufacturing among others. Most specification challenges arise from the need to provide specifications for the perceptual, reasoning and behavioural processes of robot systems that will need to acquire models of, and deal with, the high variability exhibited in human behaviour. The necessity to infer human mental states (such as beliefs and intentions) interactively, typically through sparse and/or sensor data from multimodal interfaces, imposes further challenges for the

principled specification of human factors and data-driven adaptation processes in robots operating in close proximity to humans, where safety and reliability are of particular importance.

- **AI in Healthcare**

Healthcare is a broad application domain which already enjoys the many benefits arising from the use of Artificial Intelligence (AI) and AI-enabled autonomy. This has ranged from more accurate and automated diagnostics, to a greater degree of autonomy in robot surgery, and entirely new approaches to drug discovery and design. The use of AI in medical diagnosis has advanced to an extent that in some settings, e.g. mammography screening, automated interpretation seems to match human expert performance in some trials. However, there remains a gap in test accuracy. It has been argued that the automated systems are not sufficiently specific to replace radiologist double reading in screening programmes [9]. These gaps also highlight the main specification challenges in this domain. Historically, the human expertise in this domain has not been explicitly codified, so that it can be hard to enumerate desired characteristics. It is clear that the specifications must include notions of invariance to instrument and operator variations, coverage of condition and severity level, etc. Beyond that, the 'semantics' of the biological features used to make fine determinations are subject to both ambiguity or informality, and variability across experts and systems. Moreover, the use of deep learning to achieve automated interpretation brings with it the need for explainability. This manifests itself in the challenge of guarding against 'shortcuts' [10], wherein the AI diagnostic system achieves high accuracy by exploiting irrelevant side variables instead of identifying the primary problem (e.g. radiographic COVID-19 detection using AI [10]). The specific challenge here is how to specify with respect to 'black box' models. In this regard, we can highlight the role of explainability and faithfulness of interpretation of semantics, and the role of pre-trained models in pipelines (see Fig. 1).

- **Swarm Robotics**

Swarm robotics provides an approach to the coordination of large numbers of robots, which is inspired from the observation of social insects [11]. Three desirable properties in any swarm robotics system are robustness, flexibility and scalability. The functionality of a swarm is emergent (e.g. aggregation, coherent ad hoc network, taxis, obstacle avoidance and object encapsulation [12]), and evolves based on the capabilities of the robots and the numbers of robots used. The overall behaviours of a swarm are not explicitly engineered in the system, but they are an emergent consequence of the interaction of individual agents with each other and the environment. This emergent functionality poses a challenge for specification. The properties of individual robots can be specified in a conventional manner, yet it is the emergent behaviours of the swarm that determine the performance of the system as a whole. The challenge is to develop specification approaches that specify properties at the swarm level that can be used to develop, verify and monitor swarm robotic systems.

- **Uncrewed Aerial Vehicles**

An uncrewed aerial vehicle (UAV) or drone is a type of aerial vehicle that is capable of autonomous flight without a pilot on board. UAVs are increasingly being applied in diverse applications, such as logistics services, agriculture, emergency response, and security. Specification of the operational environment of UAVs is often challenging due to the complexity and uncertainty of the environments that UAVs need to operate in. For instance, in parcel delivery using UAVs in urban environments, there can be uncertain flight conditions (e.g. wind gradients), and highly dynamic and uncertain airspace (e.g. other UAVs in operation). Recent advances in machine learning offer the potential to increase the autonomy of UAVs in uncertain environments by allowing them to learn from experience. For example, machine learning can be used to enable UAVs to learn novel manoeuvres to achieve perched landings in uncertain windy conditions [13]. In these contexts, a key challenge is how to specify how the system should deal with situations that go beyond the limits of its training (Fig. 1).

In this section, we explored what it means to specify for trustworthiness, by considering a number of ASs domains where each domain has its unique specification challenges. In automated driving, two key challenges are the lack of machine-readable specifications and the need to specify actions of other road users. In emergency and disaster settings, enhancing collaboration between autonomous agents and different human agents, and specifying human behaviour are demanding. The high variability exhibited by human behaviour is a key challenge in interactive robot systems, while in healthcare, specifying with respect to black box models is challenging. Meanwhile, in swarm robotics and UAVs, two key challenges are: specifying properties of emerging behaviour at the swarm level, and specifying how the system must handle situations that go beyond its training, respectively.

In addition, there is also a range of purely software-based ASs, i.e. systems without physical embodiment. Besides the medical diagnostic tools already mentioned earlier under AI in healthcare, such ASs are also sometimes used in justice for sentencing, in recruitment for filtering and shortlisting job applicants, and in education for proctoring, for

marking and for personalising tuition. These systems are also routinely used in finance to make recommendations on investments and manage funds. In its proposed regulations for AI, the EU [14] introduces a classification of such systems according to their risk levels. While many of these systems pose limited to no risk and can contribute to solving many societal challenges, some present risks that designers and deployers alike must address to avoid undesirable outcomes to individuals and society.

3 Intellectual Challenges for the Autonomous Systems Community

We now discuss open research questions, termed intellectual challenges, related to specifying for trustworthiness in ASs (see Fig. 2). For each challenge, we aim to identify high priority research directions.

- **How to specify human behaviour for human-AS cooperation?**

How to model human behaviour to enable cooperation with AS is challenging but crucial for the resilience of the system as a whole. It is the diversity in human enactment that drives the uncertainty about what people do and don't do, and, subsequently, the way human behaviour can be specified. Knowing the mental state of others enables ASs to steer a cooperation that is consistent with the needs of the ASs, as well as to respond to the needs of human agents in an appropriate manner.

Different theories of human behaviours explain diversity in human action in different ways and by detecting various determinants of human behaviour. For example, a behaviourist approach suggests that every behaviour is a response to a certain stimulus [15]. Albeit true, this approach is restrictive in addressing the complexity of human behaviour, as well as the different ways that human behaviour develops during cooperation. To grasp that humans are embodied with purposes and goals that affect each other, the concept of joint-action can be introduced as "a social interaction whereby two or more individuals coordinate their actions in space and time to bring about change in the environment" [16]. Adapting it to human-robot interaction, this approach is suggestive of an interplay between humans and ASs, such that what matters is not only how the AS understands the system but also how humans understand the way the autonomous agent behaves and is willing to cooperate [17]. Thus, cooperation arises from a shared understanding between agents, which is a challenge to specify.

The social identity approach [18] induces this concept of a shared understanding by providing an explanation of human behaviour focusing on how social structures act upon cognition. It proposes that, alongside our personal identity, our personality - who we are, we also have multiple social identities based on social categories and groups. Previous research has shown that social identities influence people's relation with technology [19]. Sharing a social identity initiates pro-social behaviours, such as helping behaviours in emergency situations [20]. People adapt their behaviour in line with their shared identities, which in turn, enhances resilience. Specifying social identities to enable cooperation is challenging. It requires answering questions such as: how do we represent different identities and how do we reason about them? Following the social identity approach to specify identities for human-autonomous agent cooperation requires an investigation of how to operationalise social identity, a psychological state, into software embedded within ASs.

- **How to specify data-driven adaptation processes and human factors?**

Specifying, designing, implementing, and deploying interactive robot systems that are trustworthy for use in scenarios where humans and robots collaborate in close proximity is challenging, given that safety and reliability in such scenarios are of particular importance. Example scenarios include assisting people with activities of daily living such as mobility [8] and dressing [7], rehabilitation robotics, adaptive assistance in intelligent vehicles, robot assistants in care homes and hospital environments, among others. The intellectual challenge the ASs community faces is the specification, design and implementation of trustworthy perceptual, cognitive, and behaviour generation processes that explicitly incorporate parametrizable models of human skills, beliefs, and intentions [21]. These models are necessary for interactive assistive systems since they need to decide not only how but also when to assist [22]. Given the large variability of human behaviour, the parameters of these user models need to be acquired interactively, typically from sparse and potentially noisy sensor data, a particularly challenging inverse problem. An additional challenge is introduced in the case of long-term human-robot interaction, where the assistive system needs to learn and take into consideration human developmental aspects, typically manifested in computational learning terms as model drift. As an example, consider the case of an assistive mobility device for children with disabilities [8]: as the child's perceptual, cognitive, emotional and motor skills develop over time, their requirements for the type, amount and frequency of the provided assistance will need to evolve. Similarly, when assisting an elderly person or someone recovering from surgery, the distributions of the human data that the robot sensors collect will vary not only according to the context but also over time. Depending on the human participant, and their underlying time-varying physiological

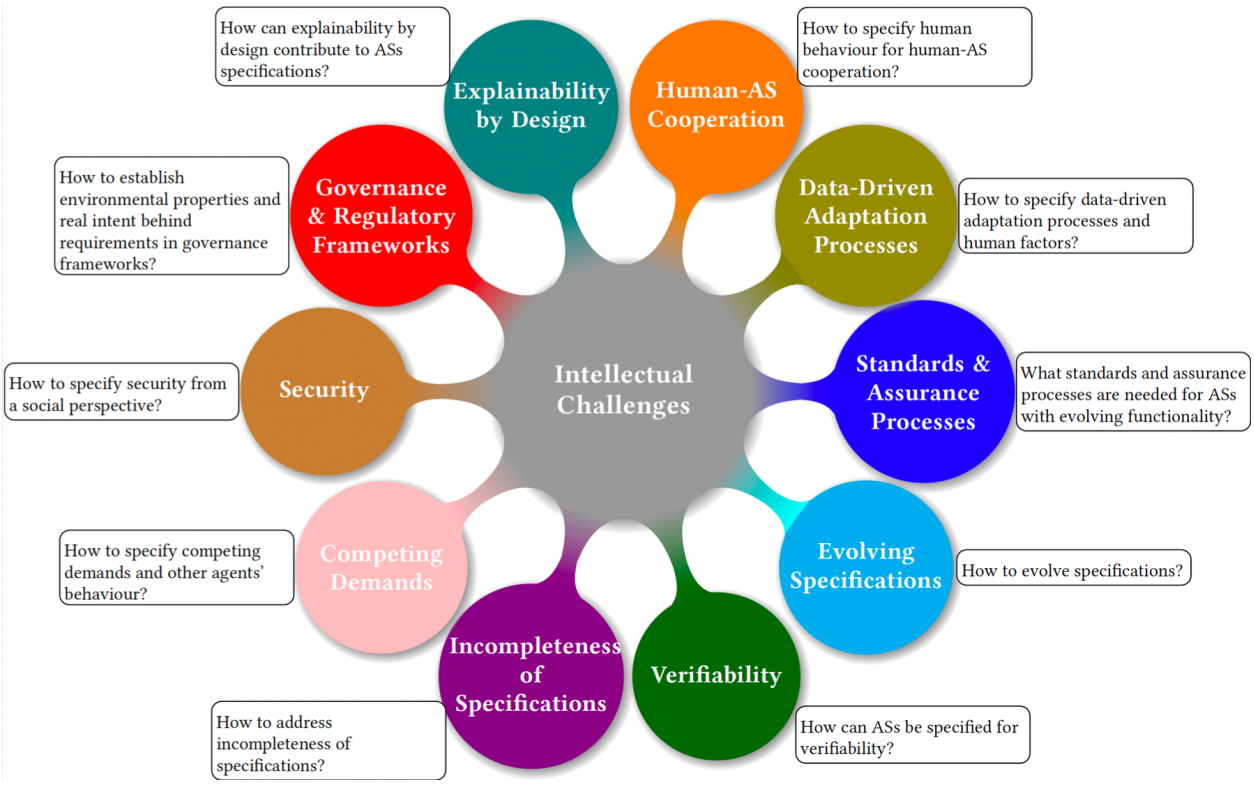

Figure 2: Intellectual challenges for the ASs community.

and behavioural particularities, model drift can be sudden, gradual, or recurring, posing significant challenges to the underlying modelling methods. Principled methods for incorporating long-term human factors into the specification, design and implementation of assistive systems, that adapt and personalise their behaviour for the benefit of their human collaborator, remain an open research challenge.

- **What standards and assurance processes are needed for ASs with evolving functionality?**

ASs with *evolving functionality*—the ability to adapt and change in function over time—pose significant challenges to current processes for specifying functionality. Most conventional processes for defining system requirements assume that these are fixed and can be defined in a complete and precise manner before the system goes into operation. Existing standards and regulations do not accommodate the adaptive nature of ASs with evolving functionality. This is a key limitation [23] that prevents promising applications such as swarm robots which adapt through emergent behaviour and UAVs with machine learning based flight control systems from deployment.

For the use of swarm robots for item storage and distribution, ISO standards have been developed for the service robotics sector (non-industrial) (e.g. ISO 13482, ISO 23482-1, ISO 23482-2), and the industrial robotics sector (e.g. ISO 10218-1, ISO 10218-2, ISO/TS 15066). Although these industry standards focus on ensuring safety of robots at the individual robot level, they do not ensure safety or any other extra-functional property at the *swarm* level that may arise through emergent behaviour.

For airborne systems and in particular for UAVs, several industry standards and regulations have been introduced to specify requirements for system design and safe operation (e.g. DO-178C, DO-254, ED279, ARP4761, NATO STANAG 4671, CAP 722). However, none of these standards or regulations cover the types of machine learning based systems which are currently being developed to enable UAVs to operate autonomously in uncertain environments.

The ability to adapt and to learn from experience are important abilities to enable ASs to operate in real world environments. When one considers the existing industry standards, they are either implicitly or explicitly based on the V&V model, which moves from requirements through design onto implementation and testing before deployment [24]. However, this model is unlikely to be suitable for systems with the ability to adapt their functionality in operation; e.g. through interaction with other agents and the environment, as is the case with swarms; or through experience-driven adaptation as is the case with machine learning. ASs with evolving functionality follow a different, much more iterative

life-cycle. Thus, there is a need for new standards and assurance processes that extend beyond design time and allow continuous certification at runtime [25].

- **How to evolve specifications?**

Every non-trivial system undergoes changes over its lifetime. The evolution of trustworthy ASs may concern changes of the requirements of their functional or non-functional properties, changes of the environment that the ASs operate in, and changes in the trust of users towards the AS. Initial specifications of the system may no longer reflect the desired properties of the system or fail to represent its actual environment.

Pursuing particular methods of evolution can make the task of specification easier. Some approaches (e.g. [26]) allow systems to evolve behaviours from sets of predefined modules. These sub-modules can be independently specified and once the behaviour is evolved these independent specifications could be aggregated to form a valid overarching specification for the system. Similarly, adaptation can be limited to a few specified behaviour states or traits which are modified reactively [27]. If configured effectively, such systems should remain within the envelope of initial specification as the system evolves or adapts within its given constraints. Designing evolvable systems in this way will not only make specifying a system an easier task but will also make the system more transparent, understandable, and potentially more trustworthy.

While much work has been done to analyse, repair, explain, or simulate specifications, little work compares specifications, detects differences, and analyses their impact on the specified system. When specifications are evolving it will be a challenge to track how they evolve, and how different they are from each other, i.e. whether the differences are purely syntactic in nature or whether there are semantic differences. In order to support this evolution of specifications, a challenge here is the semantic comparison of a new instance of the specification as it has evolved. Towards this, we may have to develop techniques to compare specifications with each other to identify differences and to understand the impact these differences may have on the system's trustworthiness.

- **How can ASs be specified for verifiability?**

For a system to be *verifiable*, a person or a tool needs to be able to check its correctness [3] with respect to its requirements and specification. The main challenge is in specifying and designing the system in such a way that this process is made as easy and intuitive as possible. For ASs in particular, specific challenges include (i) capturing and formalizing requirements including functionality, safety, security, performance and, beyond these, any additional non-functional requirements purely needed to demonstrate trustworthiness; (ii) handling flexibility, adaptation and learning; and (iii) managing the inherent complexity and heterogeneity of both the AS and the environment it operates in.

Specifications need to represent the different aspects of the overall system in a way that is natural to domain experts, facilitates modelling and analysis, provides transparency of how the AS works and gives insights into the reasons that motivate its decisions. To specify for verifiability, a specification framework will need to offer a variety of domain abstractions to represent the diverse, flexible and possibly evolving requirements ASs are expected to satisfy. Furthermore, the underlying verification framework should connect all these domain abstractions to allow an analysis of their interaction. This is a key challenge in specification for verifiability in ASs.

ASs can be distinguished using two criteria: the degree of autonomy and adaption, and the criticality of the application which can range from harmless to safety-critical. We can consider which techniques or their combinations are needed for V&V at the different stages of the system life-cycle. When ASs operate in uncontrolled environments, where there is a need for autonomy, learning and adaptation, the need for runtime V&V emerges. There, a significant challenge is finding rigorous techniques for the specification and V&V of safety-critical ASs where requirements are often vague, flexible and may contain uncertainty and fuzziness. V&V at design time can only provide a partial solution, and more research is needed to understand how best to specify and verify learning and adaptive systems by combining design time with runtime techniques. Finally, identifying the design principles that enable V&V of ASs is a key pre-requisite to promote verifiability to a first-class design goal alongside functionality, safety, security and performance.

- **How to address incompleteness of specifications?**

Incompleteness is a common property of specifications. Only the use of suitable abstractions allows for coping with the complexity of systems [28]. However, there is an important difference in the incompleteness introduced by abstractions, the process of eliminating unnecessary detail to focus e.g. on behavioural, structural, or security-related aspects of a system, and the incompleteness related to the purpose of the specification, i.e. the faithful representation of the system in an abstraction.

On the one hand, if the purpose of creating and analysing a specification is to examine a system and to learn about possible constraints, then incompleteness ("partiality" in [29]) of the representation in the specification is important as it allows for obtaining feedback with low investment in specification development, e.g. for the reduction of ambiguity [29]. On the other hand, if the purpose of the specification is to prove a property, then incompleteness of the representation may lead to incorrect analyses results. These incorrect analysis results may manifest in false positives or false negatives. False positives are often treated by adding the missing knowledge to the specification. For example, consider the specification of an infusion pump from [30] and the property that all actions are reversible. The specification reported a false positive due to an advanced handling of overflows that was missing from the specification (incompleteness). The property in the specification had to be changed to a "much more complex" one [30] to remove the false positive.

In addition to analysis tasks, specifications are also used in synthesis tasks. This is where the incompleteness of the specifications can manifest itself in the construction of biased or incorrect systems. Potentially, the synthesis of ASs from incomplete specifications may simply fail if assumptions on the environment are incomplete. As an example, consider a robot operating in a warehouse [31]. The specification requires that the robot never hit a wall. With no assumptions about the environment, the synthesizer would take the worst-case view, i.e. walls move and hit the robot, and consequently report that the specification is not realisable and no implementation exists. Adding the assumption that walls cannot move as an environment constraint changes the outcome of the synthesis. Similarly, without this assumption, a false positive result can also be expected during verification. Interestingly, when formulating requirements for humans, common sense allows us to cope with this type of incompleteness. However, the automated analysis of specifications brings with it the challenge of identifying (all) areas of incompleteness.

- **How to specify competing demands and other agents' behaviour?**

Conventional approaches to V&V for ASs may seek to attain coverage against a specification to demonstrate assurance of functionality and compliance with safety regulations or legal frameworks. Such properties may be derived from existing legal or regulatory frameworks, e.g. the UK Highway Code for driving, which can then be converted into formal expressions for automatic checking [32].

But optimal safety does not imply optimal trust, and just because an AS follows rules does not mean that it will be accepted as a trustworthy system in human society. We can also say that strictly following safety rules may even be detrimental to other properties, e.g. performance. Consider an automated vehicle trying to make progress through a busy market square full of people slowly walking across the road uncommitted to the usual observation of road conduct. The *safest* option for the AS is to wait until the route ahead is completely clear before moving on, as by taking this option you do not endanger any other road user. However, *better performance* may be to creep forward in a bid to promote your likelihood of success. Driving then, is much more than following safety rules, which makes this a particularly hard specification challenge. In this scenario an assertive driving style would make more progress than a risk-averse one.

In reality there will be significantly more considerations than just safety and performance, but this example illustrates the principle of conflicting demands between assessment standards. Consideration of other agents, such as properties of fairness or cooperation, would lead to a more trustworthy system. Additionally, the interaction of ASs with people may require insight into *social norms* of which there is no written standard by which these can be judged. Will the task of specification first require a codex of social interaction norms to be drawn together to add to the standards by which trust can be measured? Specifications would need to be written with reference to these standards, regulations and ethical principles, some of which do not currently exist, in order to ensure that any assessment captures the full spectrum of these trustworthiness criteria.

A further challenge to specification will not just be to capture the full gamut of properties for trustworthiness, but also in deciding a hierarchy through the meta-trustworthiness stack. Similar to the Asimov law of robotics, where safety to humans will outrank that of self-preservation of the robot, a similar system of hierarchy may be required. It may also be useful to consider which aspects of the specification have soft or hard limits, i.e. which are inviolable and which may be traded-off against other properties as in the case of safety versus performance or other competing demands.

- **How to specify security from a social perspective?**

There are technical sides to security, but there are also social dimensions that matter when considering how an AS enforces its status as secure. In this context, security overlaps with trust. One can only be assured a system is secure, if one trusts that system. Public trust is a complex issue, shot through with media, emotions, politics, and competing interests. How do we go about specifying security in a social sense?

On the technical side, there are fairly specific definitions for specification which can be grasped and measured. From the social perspective, the possibility of specification relies on a network of shared assumptions and beliefs that are difficult to unify. In fact, much of the value from engagement over social specifications derives from the diversity and difference. A predominant concern in social aspects of security is where data is shared between systems (social-material

interactions). That is, whenever an AS communicates with a human being or an aspect of the environment. Although these interactions have technical answers, to find answers that consider social science perspectives requires collaboration and agile methods to facilitate that collaboration.

The human dimension means that it is not enough to specify technical components. Specifications must also capture believes, desires, fears and at times misinformation with respect to how those are understood, regarded and perceived by the public. For example, in what ways can we regard pedestrians as passive users of automated vehicles? How are automated vehicles regarded by the public, and how are pedestrians involved in automated mobility?

The ethical challenges that emerge for ASs security also relate to the legal and social ones. The difficulty centres around how to create regulations and specifications on a technical level, that are also useful socially, facilitating responsiveness to new technologies that are neither simply techno-phobic nor passively accepting. Doing so must involve both innovation and public input, so that the technology developed works for everyone. The ELSI (Ethical, Legal & Social Implications) framework [33] is an example of a framework aimed to engage designers, engineers, and public bodies in answering these questions. ELSI is an inherently cross-disciplinary set of approaches for tackling ASs security, as many interrelated and entangled aspects. Specifying security requires connection, collaboration, and agile ethical methods.

- **How to establish environmental properties and real intent behind requirements in governance frameworks?**

Computer scientists treat specifications as precise objects, often derived from requirements by purging features such that they are defined with respect to environment properties that can be relied on regardless of the machine's behaviour. Emerging ASs applications in human-centred environments can challenge this way of thinking, particularly so because the environment properties may not be fully understood, or because it is hard to establish if the real intent behind a requirement can be verified. These gaps should be addressed in governance frameworks in order to engender trust.

For instance, in all of the domains mentioned above, we are increasingly seeing systems that are data-first and subject to continuous deployment. This has the interesting consequence that sometimes even the task requirements are really only given in terms of fit to observed human behaviour [34], e.g. what does an expert say in radiology interpretation and does the AI-based ASs match that [9]. Furthermore, many emerging concerns such as fairness are not only difficult to formalise in the sense of software specification, but also their many definitions can be conflicting such that it is impossible to satisfy all of them in a given system [35].

ASs of the future will need a combination of informal and formal mechanisms for governance. In domains such as automated vehicles, trustworthiness of the system may require a complete ecosystem approach [36] involving community-defined scenario libraries, enabling the greater use of simulation in verification, and independent audits via independent third parties. This calls for the development of new computational tools for performance and error characterisation, systematic adversarial testing with respect to a range of different specification types, and causal explanations that address not only a single instance of a decision but better expose informational dependencies that are useful for identifying edge cases and delineating operational design domains.

In addition to all of these technical tools, there is a need to understand the human-machine context in a more holistic manner, as this is really the target of effective governance. People's trust in an AS is not determined by technical reliability alone. Instead, the expectations of responsibility and accountability are associated with the human team involved in the design and deployment of the system, and the organisational design behind the system. A vast majority of system failures are really failures arising from mistakes made in this 'outer loop'. Therefore, effective regulations must first begin with a comprehensive mapping of responsibilities that must be governed, so that computational solutions can be tailored to address these needs. Furthermore, there is a need for ethnographic understanding of ASs being used in context, which could help focus technical effort on the real barriers to trustworthiness.

- **How can explainability by design contribute to ASs specifications?**

There are increasing calls for explainability in ASs, with emerging frameworks and guidance [37] pointing to the need for AI to provide explanations about decision making. A challenge with specifying such explainability is that existing frameworks and guidance are not prescriptive: what is an actual explanation and how should an explanation be constructed? Furthermore, frameworks and guidance tend to be concerned with AI in general, and not ASs.

A case study addressing regulatory requirements on explainability of automated decisions in the context of a loan application [38] provided foundations for a systematic approach. Within this context, explanations can act as external detective controls, as they provide specific information to justify the decision reached and help the user take corrective actions [39]. But explanations can also act as internal detective controls, i.e. a mechanism for organisations to demonstrate compliance to the regulatory frameworks they have to implement. Both [38] and [39] provide systematic questions that an explanation designer needs to consider: What is the purpose of an explanation? What is the audience

of an explanation? What is the information that it should contain? However, to adequately address these questions, explainability should not be seen as an after-thought, but as an integral part of the specification and design of a system, leading to explainability requirements to be given the same level of importance as all other aspects of a system.

In the context of trustworthy ASs, emerging ASs regulations could be used to drive the socio-technical analysis of explainability. A particular emphasis would have to be on the autonomy and the hand-over between systems and humans that characterise trustworthy ASs. The audience of explanations will also be critical, from users and consumers to businesses, organisations and regulators. Finally, considerations for post-mortem explanations, in case of crash or disaster situations involving ASs, should lead to adequate architectural design for explainability.

4 Summary and Conclusions

As ASs are becoming part of our daily lives and interacting more closely with humans, we need to build systems worthy of trust regarding safety, security and other non-functional properties. In this paper we have first examined ASs domains of different levels of maturity and then identified their specification challenges and related research directions. One of these challenges is the formalisation of knowledge easily grasped by humans so that it becomes interpretable by machines. Prominent examples include the specification of driving regulations for AVs, and the specification of human knowledge expertise in the context of AI-based medical diagnostics. How to specify and model human behaviour, intent and mental state is a further challenge that is common to all domains where humans interact closely with ASs, such as in human-robot collaborative environments as found in smart manufacturing. Alternative approaches involve the specification of norms to characterise the desired behaviour of ASs, which regulate what the system should do or should not do. An emerging direction of research is the design of monitors to observe the system and check compliance with norms [40]. The example of swarm robotics raises the need and challenge to specify behaviour that emerges at the system level and relies on certain actions of the entities that form the system with each other and their environment.

Beyond the technical aspects, across the specific ASs domains, are research challenges related to governance and regulation for trustworthiness, requiring a holistic and human centred approach to specification focused on responsibility and accountability, and enabling explainability from the outset. Fundamental to specifying for trustworthiness is a sound understanding of human behaviour and expectations, as well as the social and ethical norms applicable when humans directly interact with ASs. As for future work, an interesting extension of this paper would be to produce a classification of properties to be specified for trustworthiness under the different intellectual challenges discussed (e.g. socio-technical properties of explainability are purpose, audience, content, timing and delivery mechanism of explanations).

We conclude that specifying for trustworthiness requires advances on the technical and engineering side, informed by new insights from social sciences and humanities research. Thus, tackling this specification challenge necessitates tight collaboration of engineers, roboticists and computer scientists with experts from psychology, sociology, law, politics and economics, as well as ethics and philosophy. Most importantly, continuous engagement with regulators and the general public will be key to trustworthy ASs.

Acknowledgments

This article is a result of the fruitful discussions held at the Specifying for Trustworthiness workshop during the Trustworthy Autonomous Systems (TAS) All Hands Meeting in September 2021. The work presented in this paper has been supported by the UK Engineering and Physical Sciences Research Council (EPSRC) under the grants: [EP/V026518/1], [EP/V026607/1], [EP/V026747/1], [EP/V026763/1], [EP/V026682/1], [EP/V026801/2] and [EP/S027238/1]. Y.D. is also supported by a Royal Academy of Engineering (RAEng) Chair in Emerging Technologies.

References

[1] Pradeep K. Murukannaiah, Nirav Ajmeri, Catholijn M. Jonker, and Munindar P. Singh. New foundations of ethical multiagent systems. In *Proc. of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, page 1706–1710, Richland, SC, 2020. International Foundation for Autonomous Agents and Multiagent Systems.

[2] Mohammad Naiseh, Caitlin M. Bentley, and Sarvapali Ramchurn. Trustworthy autonomous systems (TAS): Engaging TAS experts in curriculum design. In *Proc. of the 2022 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, 2022.

[3] International Organization for Standardization. ISO/IEC/IEEE 24765:2017 Systems and software engineering — Vocabulary. Online, 2017.

[4] Hadas Kress-Gazit, Kerstin Eder, Guy Hoffman, Henny Admoni, Brenna Argall, Rüdiger Ehlers, Christoffer Heckman, Nils Jansen, Ross Knepper, Jan Křetínský, Shelly Levy-Tzedek, Jamy Li, Todd Murphey, Laurel Riek, and Dorsa Sadigh. Formalizing and guaranteeing human-robot interaction. *Commun. ACM*, 64(9):78–84, August 2021.

[5] Department for Transport. The highway code – Using the road (159 to 203). Online, 2022. GOV.UK.

[6] Keith James. The organizational science of disaster/terrorism prevention and response: Theory-building toward the future of the field. *Journal of Organizational Behavior*, 32(7):1013–1032, 2011.

[7] Yixing Gao, Hyung Jin Chang, and Yiannis Demiris. User modelling using multimodal information for personalised dressing assistance. *IEEE Access*, 8:45700–45714, 2020.

[8] Harold Soh and Yiannis Demiris. Learning assistance by demonstration: Smart mobility with shared control and paired haptic controllers. *J. Hum.-Robot Interact.*, 4(3):76–100, December 2015.

[9] Karoline Freeman, Julia Geppert, Chris Stinton, Daniel Todkill, Samantha Johnson, Aileen Clarke, and Sian Taylor-Phillips. Use of artificial intelligence for image analysis in breast cancer screening programmes: Systematic review of test accuracy. *BMJ*, 374, 2021.

[10] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, pages 1–10, 2021.

[11] Erol Şahin. Swarm robotics: From sources of inspiration to domains of application. In Erol Şahin and William M. Spears, editors, *Swarm Robotics*, pages 10–20, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[12] Alan F. T. Winfield and Julien Nembrini. Safety in numbers: Fault-tolerance in robot swarms. *International Journal on Modelling Identification and Control*, 1(1):30–37, 2006.

[13] Liam Fletcher, Robert Clarke, Tom Richardson, and Mark Hansen. Reinforcement learning for a perched landing in the presence of wind. In *AIAA SciTech 2021 Forum*. American Institution of Aeronautics and Astronautics, January 2021.

[14] European Commission. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legistative acts; com (2021) 206 final, 2021.

[15] Joe E Heimlich and Nicole M Ardoin. Understanding behavior to understand behavior change: A literature review. *Environmental education research*, 14(3):215–237, 2008.

[16] Natalie Sebanz, Harold Bekkering, and Günther Knoblich. Joint action: Bodies and minds moving together. *Trends in cognitive sciences*, 10(2):70–76, 2006.

[17] Elena Corina Grigore, Kerstin Eder, Anthony G. Pipe, Chris Melhuish, and Ute Leonards. Joint action understanding improves robot-to-human object handover. In *Proc. of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4622–4629, 2013.

[18] Russell Spears. Social influence and group identity. *Annual Review of Psychology*, 72:367–390, 2021.

[19] Younghwa Lee, Jintae Lee, and Zoonky Lee. The effect of self identity and social identity on technology acceptance. In *Proc. of the International Conference on Information Systems*, page 59, 2001.

[20] John Drury. The role of social identity processes in mass emergency behaviour: An integrative review. *European Review of Social Psychology*, 29(1):38–81, 2018.

[21] Yiannis Demiris. Prediction of intent in robotics and multi-agent systems. *Cognitive Processing*, 8(3):151–158, 2007.

[22] Theodosis Georgiou and Yiannis Demiris. Adaptive user modelling in car racing games using behavioural and physiological data. *User Modelling and User-Adapted Interaction*, 27:267–311, 2017.

[23] Michael Fisher, Viviana Mascardi, Kristin Yvonne Rozier, Bernd-Holger Schlingloff, Michael Winikoff, and Neil Yorke-Smith. Towards a framework for certification of reliable autonomous systems. *Autonomous Agents and Multi-Agent Systems*, 35(1):8, 2020.

[24] Yan Jia, John McDermid, Tom Lawton, and Ibrahim Habli. The role of explainability in assuring safety of machine learning in healthcare, 2021.

[25] John Rushby. Runtime certification. In Martin Leucker, editor, *Runtime Verification*, pages 21–35, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[26] Gianpiero Francesca, Manuele Brambilla, Arne Brutschy, Vito Trianni, and Mauro Birattari. AutoMoDe: A novel approach to the automatic design of control software for robot swarms. *Swarm Intelligence*, 8(2):89–112, 2014.

[27] James Wilson, Jon Timmis, and Andy Tyrrell. An amalgamation of hormone inspired arbitration systems for application in robot swarms. *Applied Sciences*, 9(17):3524, 2019.

[28] Jeff Kramer. Is abstraction the key to computing? *Commun. ACM*, 50(4):36–42, 2007.

[29] Daniel Jackson. Alloy: A language and tool for exploring software designs. *Commun. ACM*, 62(9):66–76, 2019.

[30] Michael D. Harrison, Paolo Masci, José Creissac Campos, and Paul Curzon. Verification of user interface software: The example of use-related safety requirements and programmable medical devices. *IEEE Trans. Hum. Mach. Syst.*, 47(6):834–846, 2017.

[31] Shahar Maoz and Jan Oliver Ringert. On the software engineering challenges of applying reactive synthesis to robotics. In Federico Ciccozzi, Davide Di Ruscio, Ivano Malavolta, Patrizio Pelliccione, and Andreas Wortmann, editors, *Proc. of the 1st International Workshop on Robotics Software Engineering, RoSE@ICSE 2018, Gothenburg, Sweden, May 28, 2018*, pages 17–22. ACM, 2018.

[32] Christopher Harper, Greg Chance, Abanoub Ghobrial, Saquib Alam, Tony Pipe, and Kerstin Eder. Safety validation of autonomous vehicles using assertion-based oracles, 2021.

[33] Maria Alejandra Luján Escalante, Monika Büscher, Katrina Petersen, Xaroula Kerasidou, Adrian Gradinar, and Hayley Alter. IsITethical? board game: Playing with speculative ethics of IT innovation in disaster and risk management. In *Proc. of the IX Latin American Conference on Human Computer Interaction*, CLIHC '19, New York, NY, USA, 2019. ACM.

[34] Eric J Topol. High-performance medicine: The convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

[35] Arvind Narayanan. Fairness definitions and their politics. In *Tutorial presented at the Conf. on Fairness, Accountability, and Transparency*, 2021.

[36] Philip Koopman, Rob Hierons, Siddartha Khastgir, John Clark, Michael Fisher, Rob Alexander, Kerstin Eder, Pete Thomas, Geoff Barrett, Philip Torr, Andrew Blake, Subramanian Ramamoorthy, and John Alexander McDermid. *Certification of highly automated vehicles for use on UK roads: Creating an industry-wide framework for safety*. Five AI Ltd, October 2019.

[37] Ronan Hamon, Henrik Junklewitz, and Ignacio Sanchez. Robustness and explainability of artificial intelligence. Technical report, Publications Office of the European Union, 2020.

[38] Trung Dong Huynh, Niko Tsakalakis, Ayah Helal, Sophie Stalla-Bourdillon, and Luc Moreau. Addressing regulatory requirements on explanations for automated decisions with provenance: A case study. *Digital Government: Research and Practice*, 2(2), January 2021.

[39] Niko Tsakalakis, Sophie Stalla-Bourdillon, Laura Carmichael, Trung Dong Huynh, Luc Moreau, and Ayah Helal. The dual function of explanations: Why it is useful to compute explanations. *Computer Law and Security Review*, 41, March 2021.

[40] Natalia Criado. Resource-bounded norm monitoring in multi-agent systems. *J. Artif. Int. Res.*, 62(1):153–192, may 2018.