On Specifying for Trustworthiness

Name of Author 1
Institute
City, Country
Email

Name of Author 2
Institute
City, Country
Email

Name of Author 3
Institute
City, Country
Email

Name of Author 4
Institute
City, Country
Email

Name of Author 5
Institute
City, Country
Email

Name of Author 6
Institute
City, Country
Email

Name of Author 7
Institute
City, Country
Email

Name of Author 8
Institute
City, Country
Email

Name of Author 9
Institute
City, Country
Email

Name of Author 10
Institute
City, Country
Email

Name of Author 11
Institute
City, Country
Email

Name of Author 12
Institute
City, Country
Email

ABSTRACT

As autonomous systems are becoming part of our daily lives, specifying for trustworthiness of these systems is crucial. ... In this article, 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. The main contribution of this article is a set of high-level intellectual challenges related to specifying a trustworthy autonomous system without focussing on how these challenges are actually realized. We also identify their potential uses in a variety of autonomous systems domains. ...

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Software and its engineering** → **Requirements analysis**; **Software organization and properties**; **Software functional properties**; **Extra-functional properties**.

KEYWORDS

autonomous systems, trust, specification

ACM Reference Format:
Name of Author 1, Name of Author 2, Name of Author 3, Name of Author 4, Name of Author 5, Name of Author 6, Name of Author 7, Name of Author 8, Name of Author 9, Name of Author 10, Name of Author 11, and Name of Author 12. 2021. On Specifying for Trustworthiness. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

An *autonomous system*is a system that involves software applications, machines and people, which is capable of taking actions with no or little human supervision [2]. Autonomous systems are no longer being confined to safety-controlled industrial settings. Instead, these systems are becoming increasingly used in our daily lives having matured across different domains, such as driverless cars, unmanned aerial vehicles, and healthcare robotics. As a result, specifying for *trustworthiness* of these systems is crucial.

According to ISO/IEC 2382:2015 standard, *specification* is a detailed formulation that provides a definitive description of a system for the purpose of developing or validating the system [1]. Writing specifications that capture trust is challenging [3]. A human trusts a robot to perform its actions, if it demonstrably acts in a safe manner. Here, the robot not only needs to be safe, but also needs to be perceived as safe by the human. In the same manner, in shared human-robot control situations, it is equally important to ensure that the robot can trust the human. This is particularly important in safety-critical scenarios when the robot needs to perform a task on behalf of the human where reasoning about trust is needed. In this regard, measuring human's current state (e.g. levels of fatigue, frustration) and assessing human's ability to actually perform the task are critical factors. To realize this, specification needs to consider formalisms that go beyond typical safety and liveness notions.

Different research disciplines define *trust* in many ways. Our study focuses on the notion that concerns the relationship between humans and autonomous systems. According to [2], autonomous systems are trustworthy depending on several key factors, such as: robustness in dynamic and uncertain environments; assurance of their design and operation; confidence they provide; explainability;

defences against attacks; governance and regulation of their design and operation; and consideration of human values and ethics.

In this article, 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. The main contribution of this article is a set of high-level intellectual challenges related to specifying a trustworthy autonomous system without focussing on how these challenges are actually realized. In this article, we discuss approaches for specifying trustworthy autonomous systems and identify their potential uses in a variety of autonomous systems domains. We conclude with a set of intellectual research challenges for the community.

2 AUTONOMOUS SYSTEMS DOMAINS AND THEIR UNIQUE CHALLENGES

Each autonomous systems domain brings about unique specification challenges:

2.1 Autonomous Driving

[Responsible Author: Subramanian Ramamoorthy]
[Source: Subramanian Ramamoorthy's presentation]

Author Guidelines: Word count: 150-300 (maximum); Format/structure: introductory sentences on the domain; specification challenge/s unique to this domain

Formal reasoning for "traffic regulation"... Analysis of traffic regulation... high-way code (e.g. when is it safe to make the turn and not to make the turn etc)...

The goal is to provide domain property analysis support for autonomous vehicles. 1) To formalise structures and logic of traffic regulations; and 2) to allow breakdown of key environment variables for vehicle operations.

High way code is written in plain English. It is written reasonably precisely. It is not formally specified as such. These kinds of code needs bit of interpretation. So how do we approach specification. Currently, we are taking Dutch traffic laws as examples, i.e. WVW and RVV (will move to the UK highway code later)...

The format of regulations defines constraints for the vehicle operations, e.g. overtaking and positioning. For each operation, there are two categories of constraints, 'must' (or 'prohibit', which is negation of 'must') and 'permit'.

Need to reason about exceptions, rule conflicts and open texture.

What are the scope of each of these rules, what are the preconditions, how do we want to deal with them, left implicit or needs further clarification in some other bit of the code. That is a specification challenge.

2.2 Emergency Responders

2.2.1 Emergency Responders. ...
[Responsible Author: Amel Bennaceur]
[Source: Amel Bennaceur's presentation]

Author Guidelines: Word count: 150-300 (maximum); Format/structure: introductory sentences on the domain; specification challenge/s unique to this domain

Emergency Responders: Why emergency scenario? Because by definition it is a disruption of usual operations. So they include a lot of unforeseen events and need for adaptation, and there is a requirement around the speed of response and adaptation of response to deal with critical situations... We need to specify functional and also SLEEC (social-legal-ethical-empathy-cultural) requirements. Those will change from one scenario to another. It is not only a design problem, there is a need for adaptation.

Amel Bennaceur described a use case using drones or autonomous agents in general to assist users... Lot of work done in assisting first responders. These are fire fighters, doctors... - 1: Interact with emergency response team to optimize the emergency response... We are interested in how to make zero responders (passers by those who don't have training) help other people... - 0: Coordinate human zero responders For that drones or autonomous agents move from simple sharing information with people into something like coordinating them, helping them collaborate. There is a very big requirement around how you communicate with those in order to ensure collaboration between them and collaboration with the autonomous agents. This actually to encourage pro-social behaviour. We want them to help. We want them to intervene in a pro-social way. Role of the autonomous agent is to coordinate the response and help them collaborate. There is also another requirement: collaborate between autonomous agents or helping the whole system (coordinate with other agents)... So to enable autonomous agents to interact with humans in these different ways is one of the things is to specify...(how to model humans to enable cooperation?)

2.2.2 Public Protection and Disaster Relief. ...
[Responsible Author: Luke Moffat]
[Source: Luke Moffat's presentation]

Author Guidelines: Word count: 150-300 (maximum); Format/structure: introductory sentences on the domain; specification challenge/s unique to this domain

Public Protection and Disaster Relief: Luke provided an example which was in the domain of public protection and disaster relief. This was a project with a consortium of developers where we essentially did creative method based workshops to try and address (specify) what potential ethical issues might be in the field when using a Pan-European 5G network for public protection disaster relief (for emergency responders). For this we used creative exercises like what you see in this Figure. During the lock down, we sent a series of exercises by post and then collaborated online. The aim of this is not just to breakdown the monotony of doing online interaction, but also to see how addressing ethical issues which can be a creative exercise. Figure - designing and facilitating online/offline remote collaboration.

2.3 Interactive Robot Systems

[Responsible Author: Yiannis Demiris]
[Source: Yiannis Demiris's presentation]

Author Guidelines: Word count: 150-300 (maximum); Format/structure: introductory sentences on the domain; specification challenge/s unique to this domain

Trustworthy Interactive Robot Systems: Robot systems that adapt their perceptual, reasoning and behavioural processes to calibrate human users' trust. Data-driven interactive systems with multiple sources of uncertainty in healthcare-related settings.

Yiannis Demiris described three examples: 1) Kids with disability learning to move around with intelligent wheel chairs; 2) Robots that interact with humans in hospitals; and 3) Robots that help one to get dressed in activities of daily living.

Figure shows "healthcare applications with close proximity: trust is a core requirement for these applications".

2.4 Healthcare Robotics

[Responsible Author: Subramanian Ramamoorthy]
[Source: Subramanian Ramamoorthy's presentation]

Author Guidelines: Word count: 150-300 (maximum); Format/structure: introductory sentences on the domain; specification challenges unique to this domain

AI-based Medical Diagnosis: Subramanian Ramamoorthy mentioned that from the ground up, it is a machine learning problem. An example: cataloging the findings - a radiologist's process...This is based on automated interpretation of radiology imaging...

What you find best models in terms of machine learning is subsymbolic neural networks-based models. What the models are looking for is essentially small cues integrated together through some sort of probabilistic fusion or other forms of integrating the information towards an overall diagnosis. This is very different from how the human experts are today being trained. A neural network is picking up cues completely different from that. So if you want to make a specification we have to first understand what is going on, and also the ways in which these systems fail (e.g. machine learning not picking up images accurately)... you need to pay attention to right cues. But you do not know what precisely to specify because all the measures that I apply in the model such as accuracy, recall and precision and sensitivity, they all can be fooled if you don't get it right...

Deep learning has revolutionised what is possible in terms of segmentation, classification and prognostics. However, there are many caveats! Usual and unprecedented forms of edge cases, e.g. did the model even look for disease in the first place?!

How to specify with respect to black box models?: 1) The design domain for operation and notions of coverage: camera properties. model features, semantics of biological features; 2) role of explainability and faithfulness of interpretation of semantics; 3) role of pre-trained models in pipelines

2.5 Swarm Robots

[Responsible Author: Dhaminda Abeywickrama]
[Source: Dhaminda Abeywickrama's presentation]

Author Guidelines: Word count: 150-300 (maximum); Format/structure: introductory sentences on the domain; specification challenge/s unique to this domain

Previous investigations have shown that users are open to adopting swarm robotic solutions, if the swarms are implemented in a trustworthy manner. For example, let us take an example of a case study that describes a public cloakroom where swarm of robots assist customers looking to deposit their jackets at an event. Specification challenges...

2.6 Unmanned Aerial Vehicles

[Responsible Author: Dhaminda Abeywickrama]
[Source: Dhaminda Abeywickrama's presentation]

Author Guidelines: Word count: 150-300 (maximum); Format/structure: introductory sentences on the domain; specification challenge/s unique to this domain

Today the operation of UAVs in applications like parcel delivery is very challenging with complex and uncertain flight conditions (such as wind gradients), and highly dynamic and uncertain air space (such as other UAVs in operation). Specification challenges....

2.7 Policy Development

[Responsible Author: Luke Moffat]
[Source: Luke Moffat's presentation]

Author Guidelines: Word count: 150-300 (maximum); Format/structure: introductory sentences on the domain; specification challenge/s unique to this domain

Project Luke Moffat works with TAS-S Node... isITethical... Ethical framework adapted to secure TAS framework where both we aim to work with engineers, designers, programmers, policy advocates and publics and non-governmental organizations. To get many voices in to the same room as possible in thinking about not just how these devices or technologies should be used when they arrive, but how these devices should be designed in the first place.. whether they are worthy of our trust and whether they are worthy of their use.

3 INTELLECTUAL CHALLENGES FOR RESEARCH COMMUNITY

Now we discuss intellectual challenges for the research community towards creating trustworthy autonomous systems...

[*Note: In order to guide the initial writing process, we have organized the different contributions to different subsections.*]

3.1 TAS Regulation and Governance

[Responsible Author: Subramanian Ramamoorthy]
[Source: Subramanian Ramamoorthy's presentation]

Author Guidelines: Word count: 300-580 (maximum)

Two issues most salient within the two case studies described previously are:

- **Environment and other agents' dynamics can be unknown, but crucial**
- **How to establish real intent behind requirements, vagueness and preferences?**

At the high-level the issue for us is the gap between formal specification way of thinking what systems got to do and how everybody else think systems should be thought of...if you take the definitions of fairness, there can be twenty valid definitions...

How and why are AI-based systems different?... we are concerned more specifically about data-driven machine learning methods, more specifically neural networks etc that are becoming the norm...

Shift from specification -> design to data -> specifications many gaps!...task is implicitly specified in data and specifications are implicit inside that data, or derived from that data....if you take the design processes that involve continuous deployment pipeline, it is challenging traditional modes of regulation...The issue in some of the models is inscrutability with respect to performance, error characteristics, etc.; contingency in any statements about behaviour.

To try to be more specific, 1) environment and other agents' dynamics can be unknown, but critical; 2) how to establish real intent behind requirements, vagueness and preferences... We are interested in open environments and open systems. There are other decision makers or entities in our environment and we have to take them into account. Typically these other environments are humans or machines operated by humans. This means we don't have very good models of what they do. We have to figure out how to specify something in the absence of this....conservative approximations do not make sense in domains like autonomous vehicles... conservative approximations can be too conservative and not very useful functionality wise. Other issue is the real intention is vague (fairness which is not vague, but it is difficult). It is hard to pin it down and when we try to pin it down, it does not capture what we try to say...these are high-level issues.

[Source: Subramanian Ramamoorthy's abstract of talk]

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 autonomous systems applications can sometimes 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. I'll try to quickly develop this idea using two use case examples - one from autonomous vehicles, and another from AI-based medical diagnostics.

3.2 TAS Functionality

[Responsible Author: Dhaminda Abeywickrama]
[Source: Dhaminda Abeywickrama's presentation]

Author Guidelines: Word count: 300-580 (maximum)

- **On the lack of industry standards on autonomous systems with emergent behaviour and learning**
- **How do you ensure safety of an autonomous system in situations where it's behaviour is an "emergent" consequence of the interaction of individual agents with each other and their environment?**
- **How do you specify an autonomous system should deal with situations that go beyond the limits of its training?**

3.3 TAS Hub

[Responsible Author: Luc Moreau]
[Source: Luc Moreau's presentation]

Author Guidelines: Word count: 300-580 (maximum)

- **How do you incorporate explainability requirements in autonomous systems designs?**

Luc Moreau said we need to think about explainability from the start when we specify and design a system. This is illustrated using the animation (plays a video)...Every day organizations use computers to make decisions that impact our lives. These could be life defining decisions like what schools our children are allocated to or whether our mortgage application is accepted or rejected. What if the computer says no? We have the right to challenge these decisions and to understand whether they were fairly made. What was the provenance of the decision (what data was used)? Where did that data come from? How was it used to make the decision? Was the decision fully automated without any human involvement? In response to growing demand for transparency and accountability around automated decision-making, PLEAD is developing the Explanation Assistance. It turns descriptions of the flow of data in a decision making system into meaningful explanations. The Explanation Assistant helps organizations demonstrate that they comply with legal regulations and makes it easier to explain complex decisions to their customers reassuring them that their data was processed as expected. More transparency improves customer satisfaction, a win-win for all. End of video.

Luc Moreau introduced the PLEAD project which is done by an interdisciplinary team. An explanation is a targeted, personalised textual or visual artifact that provides a focused description of the behaviour of the system based on data, processes, people and organisations that have influenced an outcome or a decision. Targeted: addressing a specific purpose for the individual(s) targeted; personalised: referring to circumstances of the individual(s) targeted; description: explanation describes aspects of a system's behaviour.

Why provenance?... WWW consortium defines provenance as: provenance is a record that describes the people, institutions, entities and activities, involved in producing, influencing, or delivery a piece of data or a thing in the world. W3C PROV - provides a strong foundation to derive explanations from the provenance of decisions.

PLEAD methodology overview.... Methodology steps: application, explanation requirements (policy analysis, requirements classification, minimum content determination), application data (collection of data categories, comparison to data flows), explanation plans, modelled provenance, explanations, explanation assistant.

Explanation requirements: policy analysis, requirements classification, minimum content determination... Each step is described by small card. Selected (sub-)steps: analyse policy requirements, apply classification framework, determine minimum content that we need to find in explanations, compare to data flows in a system, explanation plans, and explanation validation. This is all packaged in the Explanation Assistant we described earlier.

In conclusion, "Explainability by design" is crucial for TAS. We are writing up the methodology (from computer science and legal perspectives). We are still collecting data (interviews, costs etc). We are interested in exploring further applications.

What about TAS? What are the legal/regulatory requirements or business needs to be met? Apply PLEAD methodology. Incorporate explainability requirements in TAS designs.

[Source: Luc Moreau's abstract of talk]

Explainability of computer-based systems, and specifically autonomous systems, is a key aspect to make them trustworthy. We argue that explainability is not an afterthought, but it needs to be weaved in the application design.

In the PLEAD project (Provenance-based and Legally-driven Explanations for Automated Decisions), we have designed a socio-technical methodology to elicit explanation requirements from laws and regulatory frameworks, to specify explanations with a clear purpose targeting a specific audience, and to construct explanations programmatically by means of explanation plans and queries over provenance logged by the system being designed.

We have applied the methodology to systems providing explanations about decisions related to credit card applications and school allocations. These decisions are taken with various degrees of autonomy, in partnership with humans. Preliminary evaluations are showing positive feedback from the participants.

3.4 TAS Resilience

[Responsible Authors: Amel Bennaceur & Anastasia Kordoni]
[Source: Amel Bennaceur's presentation]

Author Guidelines: Word count: 300-580 (maximum)

- **How to specify humans to enable cooperation with autonomous systems?**
- **How to specify identities for human-autonomous system cooperation?**

In this talk, Amel Bennaceur concentrated on one specific challenge and one specific example (application) in resilience. Challenge: Corporation between autonomous systems and humans. The context (example): resilience in emergency scenarios.

On specification challenge: How to specify (how to model humans to enable cooperation)? Look at from a behaviour view (behaviourist of the user), as an automata for example. But, this is a very restrictive way of specifying how a human behaves. There is uncertainty about the actions people can do. Humans don't behave the way we expect them to, also the computers also don't behave the way we expect them to...So there is a kind of evolution to an ontological view (from behaviourist view) where you don't want to consider actions but you also want (i) opportunity the user can have (e.g. user is less than 5m away); (ii) their willingness to take certain actions; (iii) capability of each user (e.g. user can hear the robot). Here specifying little more details... But, this ontological view is still not enough, as it does not show the dynamics between different people.

So there is another way, we started to looking at which from a "Joint cognition perspective". It is not only how the autonomous system sees the system, also it is important to understand how the human sees the autonomous agent. This may help in specifications, such as explainability... Joint cognition: 1) e.g. user understand how robots behave and is willing to cooperate; 2) could it be translated into other variables (e.g. social identity markers, cooperation index)... Cooperation is not only a reaction to actions. Actually, it goes further. There is a joint cognition and there is almost a social identity shared between the autonomous agents and humans. Its is a matter of understanding what are the markers for this identity and for this com..?? To sum up, there is multiple ways for looking at cooperation between humans and autonomous agents...Behaviour is only one facet. A view only on behaviour is not enough. Need for trade off between multiple requirements... lot of questions about what are the markers we need, how do we reason about them, and how we make trade off between multiple requirements such as safety, explainability... what to build trust in autonomous systems.

[Source: Amel Bennaceur's abstract of talk]

REASON aims to expand the ability of autonomous agents to cooperate with peer agents and humans, and to proactively seek such cooperation for mutual benefit—opening up significant new opportunities for enhanced resilience through the pooling of information and functional capabilities. While significant progress has been made for specifying and enabling cooperation between autonomous systems, cooperation with humans raises additional challenges which this talk explores.

[Source: Breakout Room-2 (Amel Bennaceur, Anastasia Kordoni, Adrian Bodenmann, Kerstin Eder)]

Amel Bennaceur to summarise the outcome of their breakout room mentioned: we went deeper to corporation.. once you want to model the corporation what are the highlights? So they talked about... there is lot of social theories that specifically the identity so how to represent it, how to understand it, how to translate it to something that computer scientists understand and can implement.

3.5 TAS Trust

[Responsible Author: Yiannis Demiris]
[Source: Yiannis Demiris's presentation]

Author Guidelines: Word count: 300-580 (maximum)

- **How to incorporate human factors and data-driven adaptation processes where safety and reliability are of particular importance?**

Yiannis Demiris presented from his work through his lab which works on assistive robotics mainly, and from the Trust node he is working on as a co-investigator.

Focussed today on specification challenges specifically on: can we define a set of principles of how trust is acquired over time, how it is adapted to context, errors, the environment and the "user", and adapt systems accordingly. (Trust for user...)

Trustworthy Interactive Robot Systems: specification challenge for us here is that robot systems need to adapt to their perceptual, reasoning and behavioural processes to calibrate the human users' trust, so we don't want the humans to over trust the robots or under trust them. The particular difficulty for us is that all systems are essentially data-driven interactive systems with multiple sources of uncertainty in healthcare-related settings in terms of safety, reliability, privacy...

Specifying an assistive cognitive architecture. We don't use formal specifications... We specify the architecture using control-theoretic ensemble approach, hierarchical representations... usually through ensemble model... Specifying an assistive cognitive architecture... Ensemble model... usually we use simulation approach.. digital twin..

***Specific specification challenges for an assistive cognitive architecture:

[Sensing] Large variability in human (cognitive and emotional) processes cannot easily be inferred by sensor-observable behaviour.... Perceive user states and actions through multimodal interfaces...

[Learning] Requirement for a (sparse-data driven) user model (needs, skills, preferences...) that can change over time... Understand and predict user needs and intentions by building user models...

[Assisting] Specifying appropriate robot behaviours considering (personalised) human short-term and long-term goals... Adapt robot behaviour to improve interaction and assistance effectiveness.

[Source: Yiannis Demiris's abstract of talk]

Interactive robotic systems are a challenging domain for formal specification research. The robot systems typically rely on noisy sensory channels and need to adapt their perceptual, reasoning and behavioural responses – for example, action execution parameters or explanation modalities – to calibrate the level of trust the human users should have when interacting with them. A particular specification challenge is the principled incorporation of human factors and data-driven adaptation processes in healthcare robots operating in close proximity to humans, where safety and reliability are of particular importance.

3.6 TAS Security

[Responsible Author: Luke Moffat]
[Source: Luke Moffat's presentation]

Author Guidelines: Word count: 300-580 (maximum)

- **How to reach the middle ground between the technical possibilities and innovations related to the actual systems and social perspective?**

Main challenge in terms of specification is how to reach this middle ground between the technical possibilities and innovations related to the actual systems and the social perspective which engages with that technical landscape in a way that can be understood and digested by the publics who will end up using the technology.

Luke Moffat goes through how he has used the ELSI framework in the past with some examples.

TECHNICAL - Systems:

On the technical side, there are fairly specific definitions for specification which you all know. From my perspective, the interested ones concern is where data is shared between systems what we called as social-material interactions... that is whenever an autonomous system communicates with a human being or an aspect of the environment and these have technical answers. So there is a certain ways you can specify how those practices operate on the technical level, but to understand how they interact with the social level requires collaboration.

How do autonomous vehicle systems communicate with other networked systems, with users, and with the environment?

SOCIAL - Actors:

So as soon as you introduce human-beings as in the figure, things get complicated, now you are not just dealing with instructions that you give to a device but dealing with believes, desires and fears sometimes misinformation so he is trying to understand how autonomous vehicles (he is working with Highways England), how they are understood, regarded and perceived by publics. And the way in which pedestrians can be considered as passive users of autonomous vehicles.

How are autonomous vehicles regarded by publics, and how are pedestrians involved in automated mobility?

TECHNO-SOCIAL (approach) - Accountabilities:

What are the ethical challenges? also, the legal and social ones. So how can you have regulation but also a social space that is responsive to new technologies in a way that is neither simply techno-phobic nor passively accepting but something that involves innovation and public input. So, the technology that you receive is what works for everyone. So, how you make ethical and accountable autonomous systems?

Engagements: An answer to this is to engage with many stakeholders as possible throughout the design process... ELSI is one method of doing it. Other methods: creative methods, ethics through design...

[Source: Luke Moffat's abstract of talk]

As autonomous systems become increasingly complex and pervasive, the task of ensuring their security and trustworthiness can be aided by considering ethical, legal, and social implications (ELSI) of AS use. In this presentation, I outline some of the challenges and opportunities in combining technical specification and social specification.

Sharing some methods used by isITethical Exchange, a community platform for emergency response and disaster and risk management, I offer some possibilities for tackling the complex and unpredictable social dimensions in which AS come to operate. I advocate for participatory specification, drawing on multiple voices and perspectives to ensure that technical specifications of AS can also incorporate what is ethical, socially responsible, and legally warranted.

3.7 TAS Verifiability

[Responsible Author: Jan Ringert]
[Source: Jan Ringert's presentation]

Author Guidelines: Word count: 300-580 (maximum)

- **Specifications for verifiability**

[Source: Jan Ringert's abstract of talk]
We present the vision of the verifiability node: our vision is to enable domain experts use the most suitable abstractions and specifications for the verification task at hand. We provide a verifiability framework that connects these different abstractions and provides holistic and system-level verification results. To achieve this flexibility the verifiability node is developing a unifying framework for specification languages, their semantics, compositions, and transformations. We provide a brief overview of a systematic review of specification languages for autonomous systems and identify gaps in current literature. We conclude with a call for collaboration with other nodes, the hub, and domain experts to provide their specification requirements for our unifying framework.

3.8 Incompleteness of Specifications

[Responsible Authors: Jan Ringert & James Wilson]
[Source: Breakout Room-3 (Jan Ringert, James Wilson, Carlos Gavidia-Calderon, Luke Moffat, Dhaminda Abeywickrama)]

Author Guidelines: Word count: 300-580 (maximum)

- **Incompleteness of specifications**

James Wilson to summarise the outcome of the breakout room mentioned that we discussed conversation about human implicit knowledge and the writing specification for something autonomous like a self-driving car or autonomous car. There are things that you need to include in that you don't necessarily have to include in one for a standard car. The example Jan gave was that autonomous vehicle might not know if you don't want to hit a wall, that you cant ask that wall to move, but then you may need to move backwards or move out of the way. There are things that are fundamental to humans knowledge and human understand that you need to specify under particular conditions and a difficulty you might encounter is identifying what you might have forgotten to specify and what might be important for an autonomous system to sort of know.

During the breakout room session, James Wilson mentioned that there are specifications for lot of these systems, but the difficulty with that is, in order to obtain a specification, we need to come to a consensus...sampling correct ideas... how to create specification within unprecedented cases?...

Jan Ringert mentioned that incomplete specifications and evolving specification are two aspects which are related to each other. One needs to question whether the specifications for robots and self-autonomous cars, are these specifications would be good enough through certification? Somebody makes a partial specification just to check whether their process is the right one.

Jan Ringert mentioned that: a lot of things are unknown, not formalized and implicit in the humans...we worked on reactive synthesis... first thing we noticed is that people not writing assumptions about the environment... If you are a developer if you drive to a wall, you need to turn around or back up to get the wall away from you. If you don't specify this explicitly as an assumption for synthesis, the synthesis has to assume the wall will follow the robot. It does not know otherwise right? Many things like that, if you write a specification, many people will forget to write that walls are stationary. They think it is not required to write as it is common knowledge of developers and user, but not of the autonomous systems which try to find a solution for the task of not hitting the wall. So there is something about analyzability of specifications we need to debug them, troubleshoot them.

Carlos Gavidia-Calderon mentioned about automated synthesis.. from software engineering area he talked about a tool to automatically fix bugs..

Jan Ringert said simply going by data there might be shortcuts as Carlos mentioned. What we had was missing implicit knowledge... there can be missing stuff in data too as Carlos mentioned. Carlos mentioned if we specify something it needs to go through quality control by experts which know. James Wilson following up on that raised a counter argument – this can provide solutions which were not thought before which can be beneficial.

Jan Ringert queried how do we know what we are specifying or what we have forgotten to specify? This knowledge is very difficult to obtain. He mentioned about some program synthesis (excel flashfield...) to come up with small programs some transformation. There is a disambiguation task with user to double check your intention for a case (this is one way to fill a knowledge gap which was not specified or under specified)...

3.9 Evolution of Specifications

[Responsible Authors: Jan Ringert & James Wilson]
[Source: Breakout Room-3 (Jan Ringert, James Wilson, Carlos Gavidia-Calderon, Luke Moffat, Dhaminda Abeywickrama)]

Author Guidelines: Word count: 300-580 (maximum)

- **Evolution of specifications**

James Wilson to summarise the outcome of the breakout room mentioned the fact that you may need to complete the specification as you move along... With these evolvable systems, you start of specifying may not be the things you specify later (you may need to add additional specification). So having a specification that can adapt and evolve along the system you have that is changing is also quite crucial.

During the breakout room session, James Wilson queried that does it imply that specifications need to adapt and evolve? Jan Ringert said they need to adapt and sometimes the environment adapts, some existing assumptions are no longer valid... You change the system which has better sensors for example... He said that specifications need to be evolvable and maintainable. One thing he was looking into is how to understand the evolution? What is the impact of that to the system?. To get some kind of tool support...

James Wilson queried how do you go about getting an evolving specification? Jan Ringert mentioned about *runtime monitoring* as an example. Also, he mentioned about *learning specifications* where you start without a specification but you have observations from a system or you have a simulation of the system and you try to learn a specification from it. Learning from data but it may be incomplete, you need to generalize because otherwise it is not learning...

Carlos Gavidia-Calderon mentioned that in industry also requirements change over time, and queried whether it is radically

different in robotics/autonomous systems? Jan Ringert mentioned it is not that different..

Jan Ringert mentioned about having very different, let's say hard constraints and also constraints that can evolve.. so you can have a heterogeneous specification which you can relate to them... Luke Moffat said you know they are evolving but do not know in what direction.

3.10 Competing Demands and Other Agents' Behaviour

[Responsible Author: Greg Chance]
[Source: Breakout Room-1 (Subramanian Ramamoorthy, Yiannis Demiris, Greg Chance, Shane Windsor)]

Author Guidelines: Word count: 300-580 (maximum)

- **Specifying competing demands and other agents' behaviour**

Subramanian Ramamoorthy to summarise the outcome of their breakout room mentioned that they looked into various aspects of what do you want to do about other agents in interactions of.. in particular we were talking about what happens with autonomous vehicles. That is not the exclusive use case you can think about domestic robots and what happens with our people in the home. In both cases and many more, you have this difficulty of specification of behaviour involves other agents and what their behaviour is which is fundamentally unknown. Possibly also non-deterministic and random in some form.

During the breakout room discussion... specification challenge: competing demands/negotiation, this conflicts with rules they are working with...

Initial discussion around highway code and how to interpret this into a logical format that can be tested.

Difficult to interpret rules such as "you shouldn't cause others to slow down" which requires insight into the causal effect of your driving on other drivers around you.

Yiannis Demiris said he's be working on pedestrian problems and discussed the crowded market street problem, where there are so many people walking in the road looking at the market than any driver has a hard time making any progress.

So what do you do, just pressing ahead slowly (against the known rules of driving) pushing against the pressure of the crowd conventional to normal driving behaviour. This becomes more case of managed conflict rather than any normal driving behaviour.

There are a number of ways in which we could interpret this situation one of them being game theory and another one being formal models. There is also the consideration of social convention and how to be fair in the driving situation. An aggressive driver would make more progress in this situation above a driver that is being fair and considerate.

Greg Chance commented on how 5AI use a "3 second rule" as a performance metric for driving behaviour and that if the vehicle takes longer than three seconds to pull out at a junction or a roundabout then this will have scored badly in its performance metric.

3.11 Role of Codification in Open Environments

[Responsible Author: Shane Windsor]
[Source: Breakout Room-1 (Subramanian Ramamoorthy, Yiannis Demiris, Greg Chance, Shane Windsor)]

Author Guidelines: Word count: 300-580 (maximum)

- **Role of code/codification in open environments and how to specify them?**

Subramanian Ramamoorthy to summarise the outcome of their breakout room mentioned: the other issue that came up for us was there are some domains that are amenable to there being a code and codification and there are some that are not. So mobility is a nice thing in some sense is that there is a desire for people to follow a code and with very rare there is an exception you go through a licensing regime that make sure everyone is somewhat aware of the code. Even if they were violating it, there is reason to believe those models exist and you can make some progress with that. If you think about home robots in particular we were thinking about a Roomba vacuum cleaner and a 3 year old child, you can't make the 3 year old to do a particular thing or there is no guarantee of compliance. So what should the robot actually do in that case? It could take a very conservative stance but then that could be bad for functionality and there nothing happens. It could try to nudge? and play a game as if were but that could become more complex. So we discussed aspects of this. Certainly there is business of what do we want to do in an open environments what is the role of the code and what are the specifications in that kind of setting...

During the breakout room session, Shane Windsor mentioned there may be a case to codify the others behaviour but there is a question on how this can be achieved we cannot regulate all other people or have models to predict how they will behave. Greg said that we can only look at this situation in terms of rational agents else we end up with an Infinity problem where we postulate every possible human error that could occur. So then we come back to the issue of being overly compliant and we end up making no progress in the market problem. Subramanian Ramamoorthy discussed an issue with his Roomba vacuum cleaner and his three year old son. This autonomous agent cannot have a model of the behaviour of his three year old son and therefore it must take a conservative approach to its behaviour, always taking the less risky approach. Subramanian Ramamoorthy also mentioned a hierarchy of safety concerns where the pinnacle of this hierarchy would be something like avoid collisions at all costs and below this might be to observe red lights but not at the expense of colliding with a pedestrian. Below this we may have social conventions and an area that is more grey in terms of the interpretation of rules (trolley problem). One of the issues with pedestrians in this context is that they are poorly defined in terms of their predictable behaviour essentially being somewhat random within some loosely defined boundaries.

3.12 Trusting System after Testing/Verification/Simulation

[Responsible Author: Kerstin Eder]
[Source: Breakout Room-2 (Amel Bennaceur, Anastasia Kordoni, Adrian Bodenmann, Kerstin Eder)]

Author Guidelines: Word count: 300-580 (maximum)

- **Why is the system produced still not trusted once it is tested/verified/simulated?**

Amel Bennaceur to summarise the outcome of their breakout room mentioned: taking us back to rethinking trustworthiness, so once you test / verify / simulate why people still don't trust the system you produced. So that kind of understanding different facets of trustworthiness is also very important.

4 CONCLUSION

ACKNOWLEDGMENTS

This article is a result of the fruitful discussions at the Specifying for Trustworthiness workshop held in conjunction with the Trustworthy Autonomous Systems (TAS) All Hands Meeting. The authors thank all the speakers and fellow participants, and the TAS Hub and EPSRC for their support.

REFERENCES

[1] International Organization for Standardization. 2015. ISO/IEC 2382:2015 Information technology — Vocabulary. Online. Retrieved October 13, 2021 from https://www.iso.org/standard/63598.html
[2] UKRI Trustworthy Autonomous Systems Hub. 2020. Our definitions. Retrieved October 9, 2021 from https://www.tas.ac.uk/our-definitions/
[3] 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. 2021. Formalizing and Guaranteeing Human-Robot Interaction. *Commun. ACM* 64, 9 (Aug. 2021), 78–84. https://doi.org/10.1145/3433637