Workshop Report: Trust Me? I'm an Autonomous Machine

Workshop report authored by: Glenn McGarry (The University of Nottingham)

Workshop organisers and facilitators: David Green, Joseph Lindley, Miriam Sturdee, Zach Mason (University of Lancaster)

1 Introduction

This research is in support of the UKRI Trustworthy Autonomous Systems (TAS) Hub project. In this report we describe the findings from a workshop titled 'Trust Me? I'm and Autonomous Machine', designed to engage with experts in industry and academia in order to capture some high-level understandings of pertinent issues around trust in automated systems. The findings indicate that **trust is a 'distributed concern'** - in other words, trust is a constitution of complex relationships, multiple concerns, and stakeholder perspectives from, for example, social, legal, technical, and business sectors. These factors and the emergent themes discussed in this report will be used in the next stage of the *Trust Me? I'm an autonomous machine* research activities to represent 'expert narratives' on trust within public consultations.

1.1 Workshop overview

Briefly, the workshop involved discussions that explored the experts' perspectives on trust based on empirical accounts of their work in the AI domain. Initial break-out discussions were aimed at producing "trust maps" to map out the key issues for a given system, the actors involved, and the connections between them. Follow up discussions with the whole groups then aimed to distil the issues to "master narratives" that encapsulate several key concepts surrounding trust autonomous systems. Although the workshop activities do not map out the problem space exhaustively, these factors and their relationships are explored in the findings section.

1.2 Practicalities

The workshop involved a series of group activities and discussions that were staged and captured online using a combination of telepresence (Zoom), whiteboard (Miro) applications, and artist's sketch notes (see appendix). The workshop lasted around 3.5 hours and discussions were audio recorded and transcribed for analysis alongside the whiteboard data and sketch note renderings. Present in the workshop were 22 experts from industry and academia and the research team members from the University of Lancaster, David Green and Joseph Lindley (workshop organisers), Zach Mason (workshop facilitator), and Miriam Sturdee (sketch notes

and whiteboard artwork); and from the University of Nottingham, Glenn McGarry (workshop facilitator).

Excluding the team, 22 experts from industry and academia participated in the workshop, as listed in Table 1 showing each participant's anonymised ID, occupation, and sector.

Table 1 - Participant list

ID	Occupation	Sector
TM1	Lecturer in Intelligent Mobility	University
TM2	CTO	Industry
TM3	Professor	University
TM4	Researcher	University
TM5	Lecturer	University
TM6	UX Writer	Media
TM7	Policy Research Fellow	University
TM8	PhD student	University
TM9	Research Associate	University
TM10	Research Fellow	University
TM11	Senior Backend Engineer	Industry
TM12	Senior AI Engineer	Industry
TM13	Head of Product	Industry
TM14	PhD Student	University
TM15	Research Fellow	University
TM16	Professor of Marketing	University
TM17	PhD Researcher	University
TM18	Associate Professor	University
TM20	UX Designer	Industry
TM21	Research Associate	University
TM22	Research Associate	University

Sector representation of the participant group is broken down as follows: 1 from media; 5 from industry; and 16 from universities. For the research team, representation from industry was felt to be critical, in order to garner expert perspectives of real-world deployment of intelligent technologies. Tangible explorations such as these were felt to be preferable over abstract conversations for conveying the concerns surrounding TAS to the general public, per the aims of the following stage of research project. For this reason, individual participants from industry were allocated to each of the three break-out groups, in order ensure a balance of perspectives in each group discussion.

2 The Findings

The overarching "master narrative" resulting from the workshop shows that **trust is complex** and is firstly, constituted of several related factors linked to a system's context of use, the stakeholders involved; and secondly based on a reciprocation of understandings among stakeholders about a system's design, deployment, and limitations brought about through principles of transparency and explainability (Figure 1).

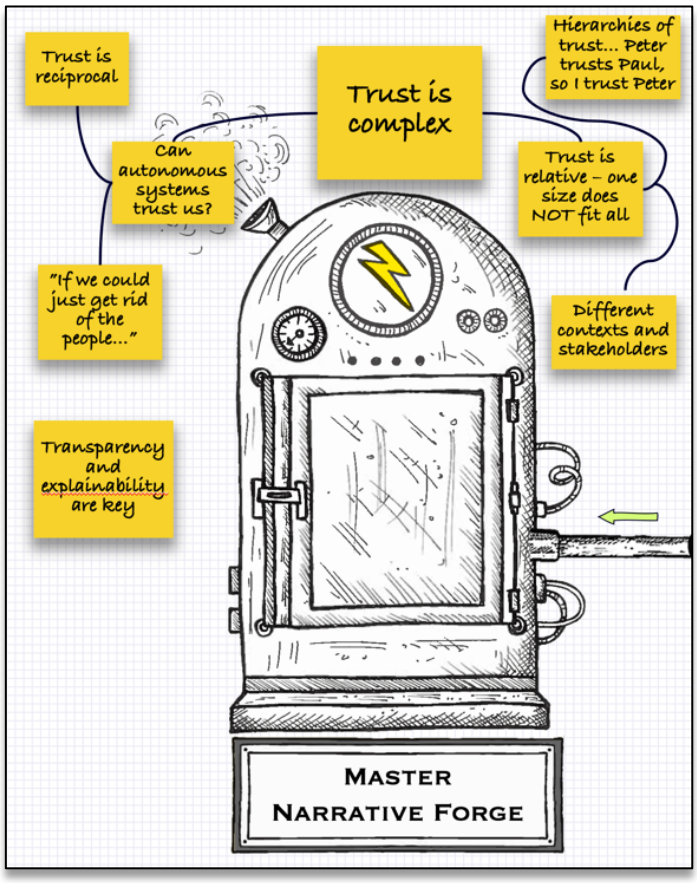


Figure 1 - The "Master Narrative Forge" – part of the illustrative artworks used to facilitate the online whiteboard sessions.

The findings presented in this report drill down from this master narrative to discuss some of the underlying narratives and themes surfaced through expert discussion, which are presented in three main sections: 2.1 Trust is Relative… one size does not fit all; 2.2 Can autonomous systems trust us?; and 2.3 Transparency and explainability is key in trust.

2.1 Trust is Relative… one size does not fit all.

This narrative proposes that different trust models hold for different circumstances for which a system is designed and deployed. Many parts of the workshop "honed-in on this idea that (trust is) very much circumstantial, and based on the application and who you're affecting, your trust is going to differ" [TM21]. In this section we explore some of the trust issues associated with autonomous systems and how they are shaped according to the context of use and the stakeholders involved.

2.1.1 Multiple stakeholders = complex trust networks

The analysis of our findings surfaces the idea that trust is a distributed concern, consisting of a multitude of related factors and trust relationships, which is exemplified in the following explanation of the stakeholders involved in a recommender system.

Dave: Who are the users of this system, and who are the other parties that are involved in the trust relationships that this autonomous system is working for?

[TM2]: First of all, there's the bit of trust (relating to the) recommendation system itself … Then there's the client platform, so the platforms that integrate (our system) recommendations, like online marketplaces, you might think of e-commerce platforms like eBay, things like that. There are also other kinds of platforms like lending platforms, gig economy platforms. Then there are the end users who actually use these platforms. There are potentially two types of end users: you have the e-commerce side; and then you have the buyers. But even then, some of them also have marketplaces themselves like Amazon, so you also have sellers. Basically, service providers, service consumers, product sellers, product buyers.

While this explanation does not detail particular trust issues, it does illustrate a complex 'ecosystem' of stakeholders and interested parties that extends beyond the direct relationship between end users and service provider to include people-to-organisation, organisation-to-organisation, and system to system trust relationships. In the example above, the expert's back-end recommendation system is syndicated across several e-commerce platforms to serve a complex 'trust network' of goods and services providers and consumers, and organisations that is facilitated through a constellation of interconnected systems. This in turn raises some significant trust concerns for systems design.

2.1.2 Design based trade-offs of trust concerns

Adding to the complexity of 'trust networks' are issues surrounding design practices that result on the trade-off of system properties related to trust concerns, and by way of example we again cite [TM2]'s example of a recommender system, as explained below.

[TM2]: We are building a system that enables people to see what their trusted friends and persons physically recommend or give feedback on stuff, so your friend so-and-so likes this auto mechanic or things like that. In that system we are building some AI components to identify the most relevant people to give feedback, and that has a lot of intricate details where things can go wrong. …

[TM2]: Depending on the level of detail that the recommendation has, if it shows you that other users prefer a certain doctor or a certain pub or whatever, then that might have privacy implications. There are various cryptographic techniques available to mitigate that, but then there is also a trade-off with respect to efficiency of the system and so on. So how can the system be trusted in this respect?

…

[TM2]: So basically, you would need to trust the system to read your situation correctly, to identify persons that you trust correctly for the particular service, and context and situation.

This example extends the notion of a 'trust network' to include people-to-people trust relationships facilitated through a system or systems, which in turn raises trust concerns specific to the systems intended purpose, and its context of use. In the case of the recommender service this relates to privacy issues, not only in terms of systematic security but in terms of trusting the system's context awareness. A particular recommendation may involve the exposure of sensitive personal information, for example the circumstances surrounding a doctor's surgery recommendation or time and location patterns that might be inferred from a taxi service recommendation. The apparent technical challenge for this use case lies in the

trade-off between these privacy issues and the efficiency of the AI part of the system to analyse and deliver personalised peer-to-peer recommendations autonomously. In more general terms, trade-offs such as these crucially turn upon the *risks* involved in relation to the system's context of use and its design constraints,

2.1.3 Trust is "the other side of the coin of risk"

In the remainder of this section, we explore some of the experts' perspectives on risk which is a theme that addresses the likelihood that an autonomous system could cause *harm* to humans, either directly or indirectly, and is an area of discussion that is key to conceptualising 'trust'.

[TM2]: "One thing that I like to think about trust is that it's sort of the other side of the coin of risk. Basically, where you require trust there is some risk that something will happen. The importance of trusting that nothing bad will happen maybe is just the same as the amount of risk that you have. If something bad happens then how bad is that really?"

A significant component of trust turns upon understanding and qualifying levels of risk, which in turn is variable depending on a system's context of use. For example, risk of harm in relation to cyber-physical systems/robotics centres on the prospect of malfunction or failure causing physical harms to people in and around its operating environment, while for cyber only systems harm manifests in different ways, for example the consequences of a personal/private data breach. For each different autonomous system and the circumstances of its use, the mix of associated risk factors may include the inherent level of risk associated with its application, for example recommender system vs automated drones; the type of risk involved, for example physical or privacy harms; and the consequences of harm. According to our experts, the importance of establishing trust is proportional to the levels of risk involved, thus safety measures for managing risk are a crucial component of trust.

2.1.4 Trust = risk managed and safe

Examples of high-risk deployment of intelligent technologies that feature prominently in the discussion involving the automated control of drone aircraft, which has applications that include emergency response, firefighting support, and surveillance. The top-level trust concern for these applications of aerial robotics, then, is safety:

[TM5]: The safety would be the first (issue) …

Glenn: Talk us through safety a little bit then, kind of elaborate.

[TM5]: There are a couple of things. One is people generally, or even the regulators, or when you talk to the CAA, they're worried about the aircraft falling from the sky for whatever reason, or crashing into a building, crashing into something it's not supposed to crash into, landing somewhere it's not supposed to land, these sort of things. So: 1) trusting the ability of the aircraft to perform what it's supposed to perform; and 2) trusting that if something goes wrong it's not going to cause a lot of damage or have a very high or significant impact on people and people's lives.

Glenn: It's how it fails then, potentially?

[TM5]: Yeah

This account of trust turns upon the safety integrity of a cyber physical system and concerns surrounding the consequences of malfunction or catastrophic failure among stakeholders that include the general public, regulatory bodies, and the system's makers and operators. For high-risk technological applications, such as autonomous drones, trust may typically be established through the system developer's proper evaluation of risk in their proposed application; the safety measures implemented around those risks; the testing of the system's functionality, reliability, and modes of failure; and the regulators arbitration of these measures against their requirements. Establishing safety is a significant trust concern that is shared across multiple stakeholder relationships and is accounted for in the kinds of risk-based governance practices detailed above. The implementation of safety measures in the design of a system, however, is not the end of the story regarding risk as in almost any technological application there is a residual risk that is inherent to its application, which raises further questions regarding the accountability for those risks.

2.1.5 Risk and accountability

Trust concerns surrounding accountability raises questions about what level of residual risk is acceptable; and who is accountable for those risks should something go wrong? These questions are particularly pertinent to the application of autonomous systems, which in current terms will almost certainly involve a 'human-in-the-loop', or a human charged with some level of operational oversight.

> [TM9]: The most advanced autonomous systems that are currently operating are at Level 4 automation, so they're almost completely autonomous but they have some level of human oversight. If you talk to people in industry and you say what about Level 5? Level 5 automation is the top one, completely autonomous and no need for human intervention, can teach themselves etc. They kind of go, "Well yeah, but we're not there yet so let's not worry about it". Even though it's probably two years maximum away. … this stuff is evolving so fast that regulation and law and stuff like that can't keep up with it.

A broadly applicable taxonomy that originates from the automotive industry defines six levels of automation from level 0 (no automation) to level 5 (fully automated systems)[1]. According to the experts, in the current landscape the state of the art in automated systems operate at level 4 (high automation), which stops short of removing human oversight completely and thus raises the question of how should accountability be divided between the parties involved?

> **[TM10]:** I think the human operator of the autonomous system (is a stakeholder in the system) as well - it depends on the level of autonomy - if we have like a human in the loop, so there is some kind of accountability for a human controller as well. Recent research actually found that when there is a human in the loop of the autonomous system, they've tried to blame the human more when mistake happens. So, these kinds of things really matter when we are designing this back-up system.
>
> …

[1] The six levels of automation adapted from SAE standards https://www.sae.org/standards/content/j3016_202104/
0) No Automation; 1) Assistive; 2) Partial; 3) Conditional; 4) High; 5) Full automation

Glenn: That's a good example, and I take your point there. Understanding what level of autonomy here we're dealing with is kind of important because if something drastically goes wrong and you're stood up in court and you're being held accountable for a part of the system that you didn't know you were accountable for, that's really important you know that from the start.

[TM10]: It is really important when we have like a high-risk task. For example, if we have autonomous social robots with the cleaning task, that's fine, there is not much thing to worry about, it's cleaning the house. But when we have some kind of healthcare robots or stuff like that, the level of autonomy I think that will have a higher impact on trust.

The problem of accountability is made up of a multitude of distributed concerns including levels of autonomy; the role of the system; the role of the human-in-the-loop; the system's context of use; the inherent/residual risks involved; the governance of novel highly automated systems; the responsibilities of the system, the system designers, and the system operators. Further to this, accountability itself presents a risk to human operators in terms of their liability should an operational failure result in harms, thus higher levels of risk and autonomy "will have a higher impact on trust" [TM10].

2.1.6 Rationalising trust as a risk/reward value: "nobody reads the terms and conditions"

In certain circumstances, trust relationships are sometimes obscured or not fully understood - such as in the transaction of terms and conditions or, as we have seen, in complex stakeholder networks – and in such cases trust may be rationalised as an ad hoc evaluation of risk/reward. Risk/reward is commonly used method of assessing the probable outcome of the circumstances in which an investment of some kind – whether this is time, effort, money, or even a medical procedure, for example – might yield a rewarding or beneficial outcome.

[TM6]: It's very much I think a risk/reward thing, because on most of these things no-one's going to read the terms and conditions, no-one's going to read how trustworthy it is or what (data) it's collecting. Facebook is a perfect example; nobody reads the terms and conditions. Facebook has gone off and allegedly done lots of interesting things with our data, and we're all still like, 'Yeah, great, I'll just carry on, or I'll come off it'. But it's still going, it's still making a tonne of money. … The next extension of that. … (is that) if there's a delivery drone that can get you your stuff in an hour but takes all your data, as opposed to a person who takes 24 hours, a lot of people are just going to go, 'I want the thing in an hour so whatever, yes please, tick all the boxes, I don't care how autonomous it is'. It's only when it crashes into your precious car or picks up your child by accident and drops it in the middle of the road … but it's only when something like that happens that people think, 'Oh, hang on a second, who's to blame?'.

By way of an example, the experts reference the social media platform Facebook and the company's historical use of their users personal data for developing and monetising their platform without these practices being explicitly apparent to their users. The catalyst for such situations is seemingly that "nobody reads the terms and conditions" [TM6] when taking up a service and "ticks all the boxes" [TM6] without fully understanding the risks involved. The practice of rationalising trust in this way is a problematic concern with potentially serious consequences; to paraphrase [TM6]'s example Facebook users risk their personal data and privacy for the reward of using a social media platform, meanwhile the end users of cyber-

physical systems, such as delivery drones, risk physical harms to themselves or their property for which they may have unknowingly accepted liability in exchange for the benefits of a faster service.

2.2 Can Autonomous systems trust us?

In this section we explore the relationships between humans and autonomous systems as actors in the social domain. These relationships and their associated challenges are often conceptualised as 'the-human-in-the-loop', which considers the role of humans in the operation of autonomous systems and in some cases, as we will discuss, the design challenges and potential for adverse effects that this entails.

2.2.1 Human feedback can improve autonomous systems… but humans don't always know best

As we have discussed so far, in the current landscape fully autonomous systems have not yet been realised and that human input at some level is often required during operation and according to the experts, feedback from the 'human-in-the-loop' is a key consideration in the design of a system's functionality that can also be exploited to improve adaptive AI systems.

> [TM14]: There are many points in favour of (a human-in-the-loop). For example, maybe the user could give an extra input … if the drones are not able to see a person but a (human) user can see, for example, I need (the drone) to go there because of this. The human should be able to provide to the system the right level of knowledge as another input. It could be as another input of the observations of the system could be considered as a human input. … there are ways that a human can improve the autonomous systems if they provide some kind of feedback.
>
> …
>
> [TM6]: I think it's really interesting that human in the loop thing, because it assumes that a human knows best, and we all know humans haven't always made the best decisions all the time. So yeah, I just find it interesting that we think that humans can save computers and AI.
>
> [TM14]: I was just talking about the human in the loop, that there are different approaches when the system has doubts about the human if the human is suggesting a wrong decision. So, there should be communication saying 'do you think that is the best thing to do or not?'; 'this is why I think it is not the right thing to do'; or, 'okay I accept your recommendation'. It's interesting.

This perspective of trust questions the trustworthiness of decision-making capabilities, not only of AI but also the human overseer of the systems' operation, in this case the real-time operation of automated surveillance drones. This in turn proposes a concept of the human-in-the-loop that recognises the inherent fallibility of both human and machine that surfaces a design challenge for a 'symbiotic' approach to decision making: the human no longer provides the manual override where the system's capabilities end but is part of a negotiation with the system that exploits each other's higher capabilities in order to make better decisions. This approach, however, also raises concerns about trusting the integrity of human decisions, as [TM6] puts it "humans haven't always made the best decisions all the time". For the case discussed above, the design challenge might be to implement controls in form of human computer dialogues

around decision making, however in other contexts trust concerns surrounding human actors manifest in other ways.

2.2.2 Trust in data driven technologies: gaming the system, data provenance, bad actors, and data quality

Part of the fundamental premise of automated systems is that they are largely data driven, whether this be the training and test data used to prime machine learning for a task; sensor data about a machine's operating environment; or users' personal data/user generated data that are exploited for personalised services. Corruption of these data in some can present a vulnerability from adversarial attack, which is a trust issue that is conventionally dealt with through security measures, however there are other ways in which a system might be corrupted in order to undermine trust. For example, this might include 'gaming' the system to create false or misleading outcomes with potentially bad consequences, which might be brought about via bad actors feeding the system with false data. This particular problem is highlighted in the example below, which again focusses on a recommender service.

[TM2]: Definitely hacking is an issue. Then there's also the issue of giving fake reviews, like you're recommending your friend's car service even though you know your friend is actually not such a great mechanic. So, for that in our system we basically allow only to give reviews when the reviewer actually has used the service, probably. Okay, probably is a big word, but if the system believes it.

…

Dave: Can the system be gamified in any way?

[TM2]: Well, yes. We tried to design it (out) as best as we could, but okay it's not proven that it is really, and therefore there will be some loopholes here and there. That's part of the challenge to design the mechanism in the way that prevents that at least for the most part. Generally, the idea is if we have good service providers and people who truthfully report their experience in terms of feedback, then all-in-all everybody should be happy in the end, except maybe the cheaters so to speak. So, if you have a bad service provider who has bad service then they might do that. That is of course a thing that is a bit difficult to sell sometimes. It's definitely not in the interests of people who can expect to get bad feedback in that system, for example.

In the example above, the 'trust model' is built on a notion 'transitive trust', which as our expert explained "is a bit trust by proxy saying okay, because you trust this person for this particular service and this context to give you a recommendation, that means that if that person recommends the service and trusts the service basically then you will also trust the service" [TM2]. In this context, the trust model effectively aims to defer trustworthiness from the system to the community of end users, however, as discussed above, the system can be gamed in order to manipulate outcomes somehow. In order to meet this design challenge, the recommender system is built around a "token economics" that rewards users who contribute reliable reviews and penalises untrustworthy users, thus the system facilitates control measures surrounding trust issues according to the system's context of use.

Trust issues surrounding data driven systems extend beyond data provenance and bad actors to a much more general problem of data quality and again we turn to the example of a recommender system

[TM2]: First of all, any system that tries to learn users' preferences like we are trying to do here to some extent, has the problem of only seeing more or less a narrow view of that user. You only know that user from interactions with the kind of hubs that you're actually integrating with and that you are getting data from. ... The system focuses or hones-in one particular aspect of that user because that's the data they got, but that might not be very well represented.

…

[TM2]: There are lot of challenges in that. First of all, of course, is how do we get to learn enough about these users to get a reasonable picture of what might be a good recommendation here. If the system only learns from certain use case data about a user, then they might get a very limited impression of that user. So that's one challenge.

…

[TM2]: Another, is that by our very mechanism we are sort of potentially building or increasing filter bubbles. Like you have groups of people who trust each other, and the system learns that, and then you only ever see recommendations from those people, you never look outside. So that makes it hard for newcomers or for maybe minorities in some cases to get a foothold in certain services or industries. There needs to be some mitigation mechanism.

Here the issues centre around the generation of 'filter bubbles' wherein the system, when given a very narrow view of a user's profile, may return equally limited recommendations and services, which is particularly problematic for context sensitive technologies such as those discussed above. The narrowness of data for general AI application then is problematic for the quality of its functionality and one which is assumed here to centre on challenges of complex user data generated through the constellations of devices services and the data privacy that surrounds those. Other issues also arise regarding filter bubbles that there is the potential for a system to create limited networks of users and services, which in turn may create unfairness – in this example at the cost of excluding some people from marketing opportunities within an online community. Careful consideration is therefore required firstly, around the provenance and quality of the data that data driven systems utilise; and secondly, to the controls that are placed around adverse external influences to mitigate the potentially bad effects of emergent properties of a system.

2.2.3 Limitations of the system's responsibilities

The cumulation of trust concerns discussed in this section so far centre on human behaviours and responsibilities associated with automated systems that include meeting the design challenges to mitigate their unwanted effects. These mitigating control measures form the notion of "the system's responsibilities" which is one part of a trust relationship that is concerned with the tensions between responsibilities of the system and the individual. Again, we turn to the example of the recommender system.

[TM2]: Ultimately our system has the premise that we show (the user) feedback from people that the user trusts, so we are trying to remove the system from the responsibility (of giving a) definitive answer. We don't say “go to this doctor”, we just say “your friends also trust this doctor”, and whether you then actually trust your friend to give you a recommendation about that is still up to the user. But of course, the system has already filtered (recommendations) from a number of possible people … (so) you are showing something to the user and not something else, so I think there's still some responsibility here of the system to give a good recommendation.

…

[TM2]: Then of course what happens if we actually give a wrong recommendation? We show you the recommendation of the person that you trust to recommend a doctor and it turns out to be a very bad decision. Then we can say that it's not the system who recommended the doctor, it's that person you know, and the system only identifies that person as a recommender, and at the end of the day it's the responsibility of the user whether they want to trust this information or not. But maybe that's a bit cheap excuse and the system actually should be held more responsible here for potential consequences. … I think it's a bit tricky here to say where exactly the responsibility of the system ends. Just because we say maybe do not trust the system, we just give you some information.

The tensions surrounding the trust relationship between individuals and the system are writ large in the example above - while the premise of the recommender system defers trust to individuals, the system still facilitates the exchange of information and therefore should be trusted to only deliver reliable and relevant information. This problem is comparable to ongoing issues surrounding social media and other user generated content platforms who have historically denied responsibility for content published on their platforms[2]. For example, according to [TM2] Facebook strategically claims that their platform “just shows you this news, but do not claim any responsibility whether that news is actually true or fake news (but) that doesn't completely remove the responsibility from the system because … we know that people will click on stuff that you show them sometimes. ... Even if the manufacturer claims that a system is not responsible, it doesn't mean that it really isn't, and where exactly the line is drawn is difficult, I would say” [TM2].

2.3 Transparency and explainability is key in trust

The third of the master narratives relates to different perspectives on transparency and explainability, which are closely related concerns that are prevalent in the academic literature on AI and are thought to be a significant factor in achieving trustworthiness in autonomous systems. While there is no single definition of these concepts, in broad terms “Transparency can be considered as the property that makes it possible to discover how and why the system made a particular decision or acted the way it did, taking into account its environment”; while

[2] https://www.theguardian.com/technology/2018/jul/02/facebook-mark-zuckerberg-platform-publisher-lawsuit
https://www.theguardian.com/technology/2020/may/28/zuckerberg-facebook-police-online-speech-trump
https://www.bbc.co.uk/news/blogs-trending-53228343

explainability "is concerned with the ability to provide explanations about the mechanisms and decisions of AI systems", particularly where something may have gone wrong [3].

The key challenges surrounding these concerns are firstly that transparency is related to a range of ethical issues surrounding AI that may differ depending on the context of use; and secondly, explainability is a considerably difficult to achieve for certain branches of AI such as machine learning, which is highly complex and difficult to reverse engineer in order to explain decisions retrospectively once the AI has begun its 'learning' processes.

Transparency and explainability is frequently addressed in research literature through proposals for ethical and technical standards, policies, frameworks, and requirements for design. The workshop discussion, however, took a different approach and offered some perspectives how these issues manifest in different contexts and how they are addressed in design practice.

2.3.1 Building in transparency through design

Achieving transparency around a system's decision making is predominantly a design challenge that requires careful consideration of the system's context of use, the stakeholders involved, and methods of communicating decisions. The example below summarises how trust issues are addresses by designing transparency into the system, in this case a heating management system for a residential apartment complex.

> [TM20]: We were looking at a heating management system. Some of the issues we looked at was how to deal with conflicting requirements. What happens if waste (energy) is configured wrong? What happens if residents aren't able to control the system? … For example, your place in the building, if everyone's getting the same sort of heating but you're feeling colder because of your location, what could that do? … We looked at transparency on decisions, so trying to explain how the system actually came up with the result it did to both occupants and also the management company. Making sure that that's in familiar clear terms using simple language.
>
> …
>
> [TM20]: What we discussed as well in the group, involving residents or a number of residents in the design process of a system like this, so you hear their points of view and they're able to in some way contribute to its development, and possibly also understand the reasoning behind it.

Design challenges often require the creation of complex systems that are required to make different decisions and for different reasons while giving due consideration to other ethical issues that are sometimes interrelated with transparency, such as fairness which is highlighted in the example above. This may at times require the deployment of appropriate design methods, such as co-design or participatory design, in order to capture what a community's sense of fairness might be – for example an apartment's acceptably comfortable temperature range – in order to incorporate those as design requirements and accordingly communicate the decision making process clearly.

---

[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8351056/

2.3.2 Building in explainability through design

As we have discussed so far, transparency and explainability, while closely related concepts, are distinct from each other in that transparency is concerned with the legibility of a system's outcomes; and explainability is concerned with understanding the underlying technological workings. The example below focusses on the explainability of a system, again related to drone aircraft, and how this is built in at the modelling phase of the design process.

[TM14]: We are trying at this stage simulator drones that provide communications in case of disasters. They move autonomously. They are self-adaptive systems that move around and try to … provide (mobile phone) coverage if the base stations are broken, or we just need extra base stations to provide more coverage.

Dave: Interesting.

[TM14]: The key thing here is that as the drones communicate each other, it's a multi-agent system, they have to trust in what the other one is doing, and also to the managers per se, to the people, for example the communication provider, to show that the drones are doing what they are meant to be doing, that they are taking the right decisions, going this way. For example, one drone is moving to the north because there is a group of people that are there, they are not covering enough space to provide the communication.

What I'm trying to do is to try to explain the history of the decision making of the drones. We're trying to explain to the developers basically what the system is doing. So, we are not explaining something for the end user, we are trying to explain it to the developers that the system works as it's supposed to be working.

Dave: Sounds fascinating. Where are you at with the development of this process? Is this a prototype at the moment, or is this something you've field tested?

[TM14]: Right now, we are just working as a simulation. The drones use machine learning for the decision making. We have tried with different algorithms, but so far, it's just a simulation.

The predominant perspective on explainability is concerned with understandings of automated system decision making in operation and often in retrospective terms and in the expert domain, for example through recordings of telemetry data from aircraft or cars. The example above, however, highlights the utility of explainability much earlier in the lifecycle of an automated system during its development stage. Here the AI component of the drone system is tested through simulation, which in turn makes the explanations of its decision making available to all the stakeholders involved in the development process which effectively 'builds-in' explainability into the system.

2.3.3 Transparency of system's intention of use and the actors involved

Finally, a different perspective on transparency that the discussions surfaced is a departure from the conventional notions and relates to public understandings of who the actors are behind a system, what their motivations are, and the intended use of the system as exemplified in the extract below on the topic of drone aircraft.

[TM5]: A lot of people are worried about if these drones are flying round and taking photos, or they have cameras on-board, who's looking at these cameras, who's looking at these photos? Even though

many of the images are only used there and then and not recorded, formatted, particularly when the aircraft is just using it for navigation, so it's not the main application to go round and take videos or images. It has a camera sensor, or some have camera sensors on-board, to aid in navigation and obstacle avoidance, and these usually just take a frame, process it, figure out if there is an obstacle or not, and then delete it. But people don't have trust in that process actually happening, they don't believe that this footage is not being kept or not being used for malicious purposes.

[TM10]: I think this brings another thing which is transparency and controllability where when you have security people who want to know, for example, where their photos are stored, this process. Also, people want to know that these ones are controllable by another human being and some degree of automation. So, it's not like throwing around on the air, they are just doing their job. They want to know that if something wrong happens there's another human controlling that as well.

The trust issues in this example relate to transparency from a public perception about the intended use of drones and the level of automation involved in their operation, which is essentially an issue of trust in the people behind the system and its intended use. The example also highlights the ambiguities that might exist around applications in the absence of transparency: on-board camera sensors are used for sensing and navigating around the drone's operating environment, but might equally be perceived as a threat to privacy without any knowledge about how images are processed, whether they are deleted or retained and stored, and by who; also, the who is in control of the drones should something go wrong? This notion returns us to issues of risk and accountability and gives a view of transparency through the lens of governance rather than of design.

3 Conclusion

The workshop findings have provided insights into the prevalent issues surrounding trust in autonomous systems from several *practical* viewpoints, including the implementation and operation of cyber-physical systems (drone aircraft); cyber-only systems (online community recommenders); and the automation of utilities (heating management). While the workshop did not explore the problem space exhaustively, the examples that were explored through expert discussion opened up and added to the empirical understandings of an array of trust issues.

To conclude, we ask the question what have we learned about trust in autonomous systems? As stated in the introduction of this report, our primary observation is that **trust is a distributed concern** that is constituted of multiple factors including the stakeholders involved, the context of use, and technologies that are deployed. For any given circumstance these factors interact in different ways that in turn give way to different perspectives on notional concerns are associated with trust in autonomous systems.

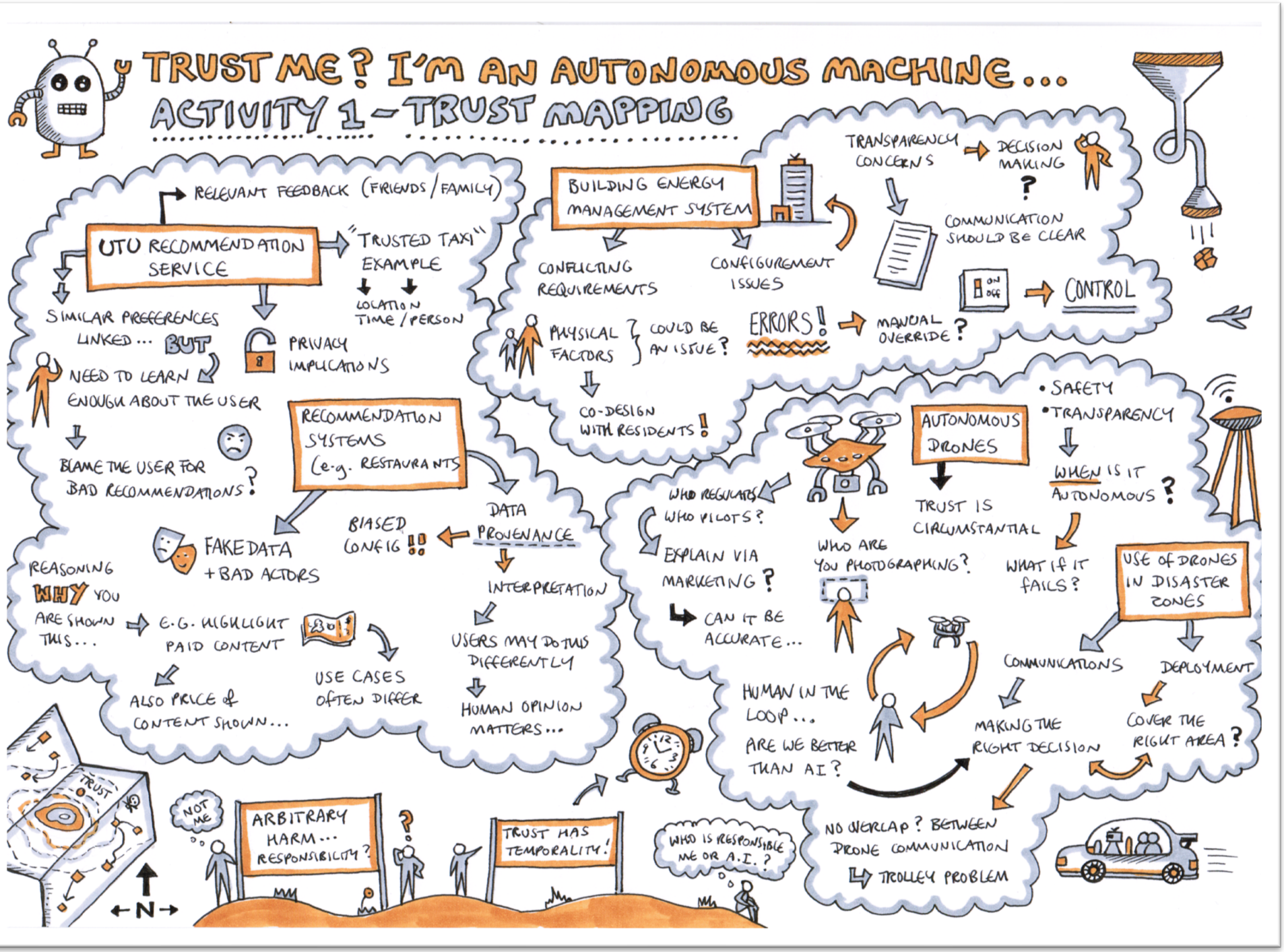
The concerns surrounding trust are often discussed in abstract terms, for example, fairness, accountability, transparency, explainability, privacy, and ethics to be resolved through frameworks, standards, or governance. While our findings do encompass these notions, they

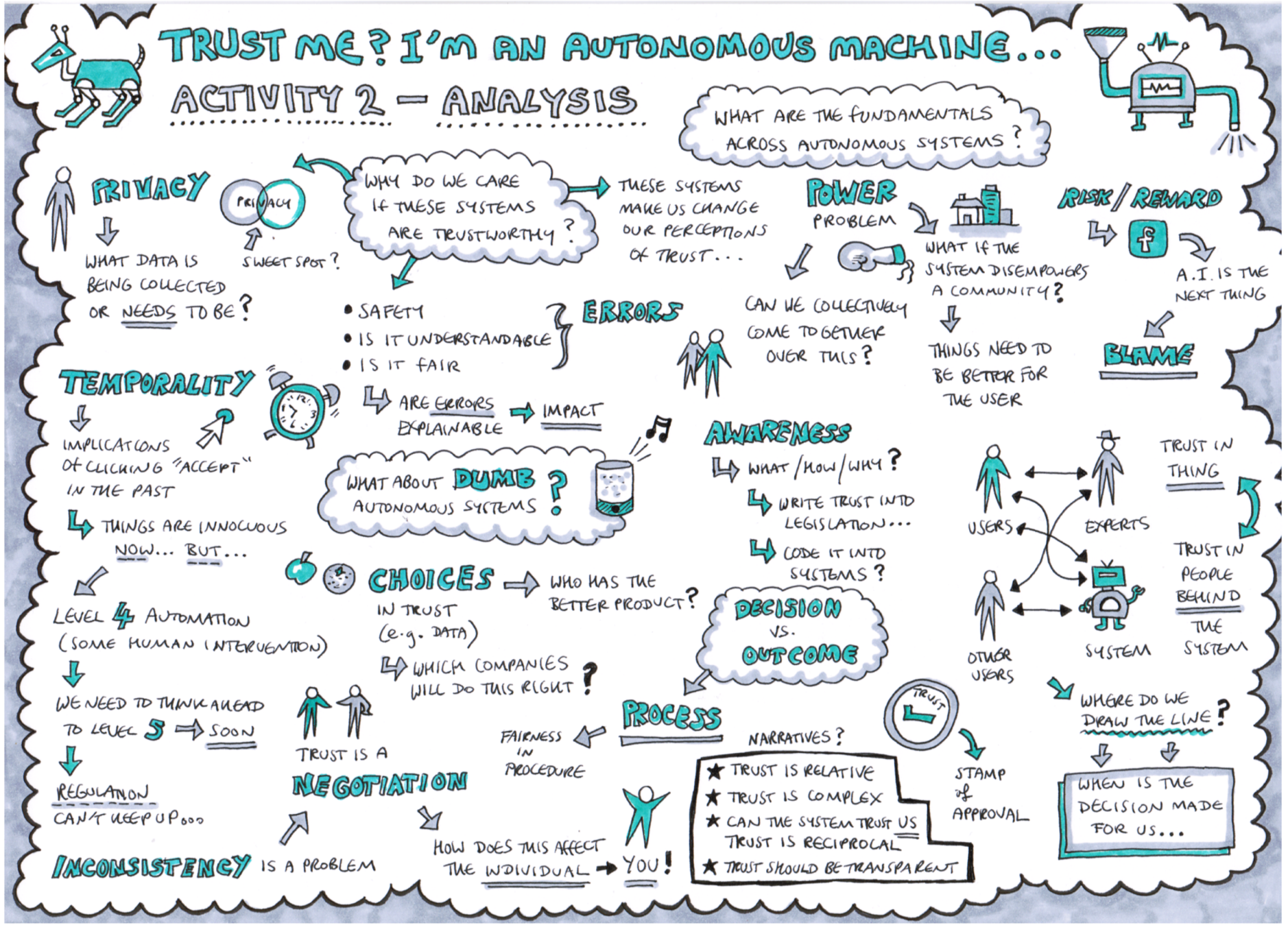
show that when autonomous systems are realised, either conceptually, through design, or proof-of-concept, these and other concerns are tricky to disaggregate as stand-alone topics within a system's context of use. The ways in which trust factors and concerns interact are subject to the circumstances that surround them, often placing them in a hierarchy or even in conflict. This is borne out in the trade-offs in system properties that were surfaced through discussion, for example, security versus reliability; or in the levels of automation and thus the accountability of the system versus the human-in-the-loop.

Underpinning these trust factors is the notion of **risk**, which can be viewed as an ethical, social or governance concern depending on the circumstances and stakeholder perspectives involved. In ethical terms, risk is addressed through transparency and explainability measures that aim to make systems understandable and accountable by design, but also to make the intentions and responsibilities of the systems' makers and operators clear and accountable. In terms of governance, regulatory bodies are typically arbitrators of a system's safety integrity and take a risk-based approach to assessing any residual harms from system failures, including catastrophic failures. Risk levels may be demonstrable through the manufacturers' safety measures implemented around those risks and the testing of the technology's functionality and reliability. Finally, in terms of social perspective of risk, the untrustworthy behaviours of people, systems, and organisations has changed the trust landscape over time. For example, "ticking the boxes" to accept terms and conditions without fully apprehending their consequences is perceived as riskier today than at the advent of, say, social media platforms, due to a slew of pre-GDPR watershed of privacy concerns. Nonetheless, as automated systems advance, their necessity may drive future changes in the trust landscape and normalise society's approach to risk as it has in the past: as expert [TM21] shared in the following anecdote about the advent of automated elevators: "nobody wanted to be in an elevator unless there was operator in it … only when the elevator operators went on strike … people saw 50 storey buildings in the middle of New York and thought, 'Actually I don't really fancy walking that'. So maybe there is an element of need and availability that affects who has what say and whose say matters".

Appendix

Workshop Sketch notes & poster submitted to TAS AHM 2021


TRUST ME? I'M AN AUTONOMOUS MACHINE...
ACTIVITY 1 - TRUST MAPPING
UTU RECOMMENDATION SERVICE
RELEVANT FEEDBACK (FRIENDS / FAMILY)
"TRUSTED TAXI" EXAMPLE
LOCATION
TIME / PERSON
SIMILAR PREFERENCES LINKED... BUT
NEED TO LEARN ENOUGH ABOUT THE USER
PRIVACY IMPLICATIONS
BLAME THE USER FOR BAD RECOMMENDATIONS?
RECOMMENDATION SYSTEMS (e.g. RESTAURANTS
FAKE DATA + BAD ACTORS
BIASED CONFIG!!
DATA PROVENANCE
REASONING WHY YOU ARE SHOWN THIS...
E.G. HIGHLIGHT PAID CONTENT
ALSO PRICE & CONTENT SHOWN...
USE CASES OFTEN DIFFER
INTERPRETATION
USERS MAY DO THIS DIFFERENTLY
HUMAN OPINION MATTERS...
BUILDING ENERGY MANAGEMENT SYSTEM
CONFLICTING REQUIREMENTS
CONFIGUREMENT ISSUES
PHYSICAL FACTORS
COULD BE AN ISSUE?
CO-DESIGN WITH RESIDENTS!
TRANSPARENCY CONCERNS
DECISION MAKING
?
COMMUNICATION SHOULD BE CLEAR
ON OFF
CONTROL
ERRORS!
MANUAL OVERRIDE?
AUTONOMOUS DRONES
WHO REGULATES WHO PILOTS?
EXPLAIN VIA MARKETING?
CAN IT BE ACCURATE...
WHO ARE YOU PHOTOGRAPHING?
TRUST IS CIRCUMSTANTIAL
WHAT IF IT FAILS?
• SAFETY
• TRANSPARENCY
WHEN IS IT AUTONOMOUS?
USE OF DRONES IN DISASTER ZONES
COMMUNICATIONS
DEPLOYMENT
HUMAN IN THE LOOP... ARE WE BETTER THAN AI?
MAKING THE RIGHT DECISION
COVER THE RIGHT AREA?
NO OVERLAP? BETWEEN DRONE COMMUNICATION
TROLLEY PROBLEM
TRUST
N
NOT ME
ARBITRARY HARM... RESPONSIBILITY?
TRUST HAS TEMPORALITY!
WHO IS RESPONSIBLE ME OR A.I.?


TRUST ME? I'M AN AUTONOMOUS MACHINE...
ACTIVITY 2 – ANALYSIS
WHAT ARE THE FUNDAMENTALS ACROSS AUTONOMOUS SYSTEMS?
PRIVACY
WHAT DATA IS BEING COLLECTED OR NEEDS TO BE?
PRIVACY
SWEET SPOT?
WHY DO WE CARE IF THESE SYSTEMS ARE TRUSTWORTHY?
THESE SYSTEMS MAKE US CHANGE OUR PERCEPTIONS OF TRUST...
POWER
PROBLEM
WHAT IF THE SYSTEM DISEMPOWERS A COMMUNITY?
CAN WE COLLECTIVELY COME TOGETHER OVER THIS?
THINGS NEED TO BE BETTER FOR THE USER
RISK / REWARD
A.I. IS THE NEXT THING
BLAME
ERRORS
SAFETY
IS IT UNDERSTANDABLE
IS IT FAIR
ARE ERRORS EXPLAINABLE → IMPACT
TEMPORALITY
IMPLICATIONS OF CLICKING "ACCEPT" IN THE PAST
THINGS ARE INNOCUOUS NOW... BUT...
WHAT ABOUT DUMB AUTONOMOUS SYSTEMS?
AWARENESS
WHAT / HOW / WHY?
WRITE TRUST INTO LEGISLATION...
CODE IT INTO SYSTEMS?
USERS
EXPERTS
OTHER USERS
SYSTEM
TRUST IN THING
TRUST IN PEOPLE BEHIND THE SYSTEM
CHOICES
IN TRUST (e.g. DATA)
WHICH COMPANIES WILL DO THIS RIGHT?
WHO HAS THE BETTER PRODUCT?
DECISION VS. OUTCOME
LEVEL 4 AUTOMATION (SOME HUMAN INTERVENTION)
WE NEED TO THINK AHEAD TO LEVEL 5 → SOON
REGULATION CAN'T KEEP UP...
TRUST IS A
NEGOTIATION
PROCESS
FAIRNESS IN PROCEDURE
NARRATIVES?
TRUST IS RELATIVE
TRUST IS COMPLEX
CAN THE SYSTEM TRUST US
TRUST IS RECIPROCAL
TRUST SHOULD BE TRANSPARENT
TRUST
STAMP of APPROVAL
WHERE DO WE DRAW THE LINE?
WHEN IS THE DECISION MADE FOR US...
INCONSISTENCY IS A PROBLEM
HOW DOES THIS AFFECT THE INDIVIDUAL → YOU!

Trust Me?... I'm an Autonomous Machine
Identifying Expert Narratives Around Trustworthy Autonomous Systems

Lancaster University | The University of Nottingham

Workshop Organisers: **David Green, Joseph Lindley, Glenn McGarry**
Artwork and Sketch Noting: **Miriam Sturdee**
Facilitation Support: **Zach Mason**

Engineering and Physical Sciences Research Council

UKRI Trustworthy Autonomous Systems Hub

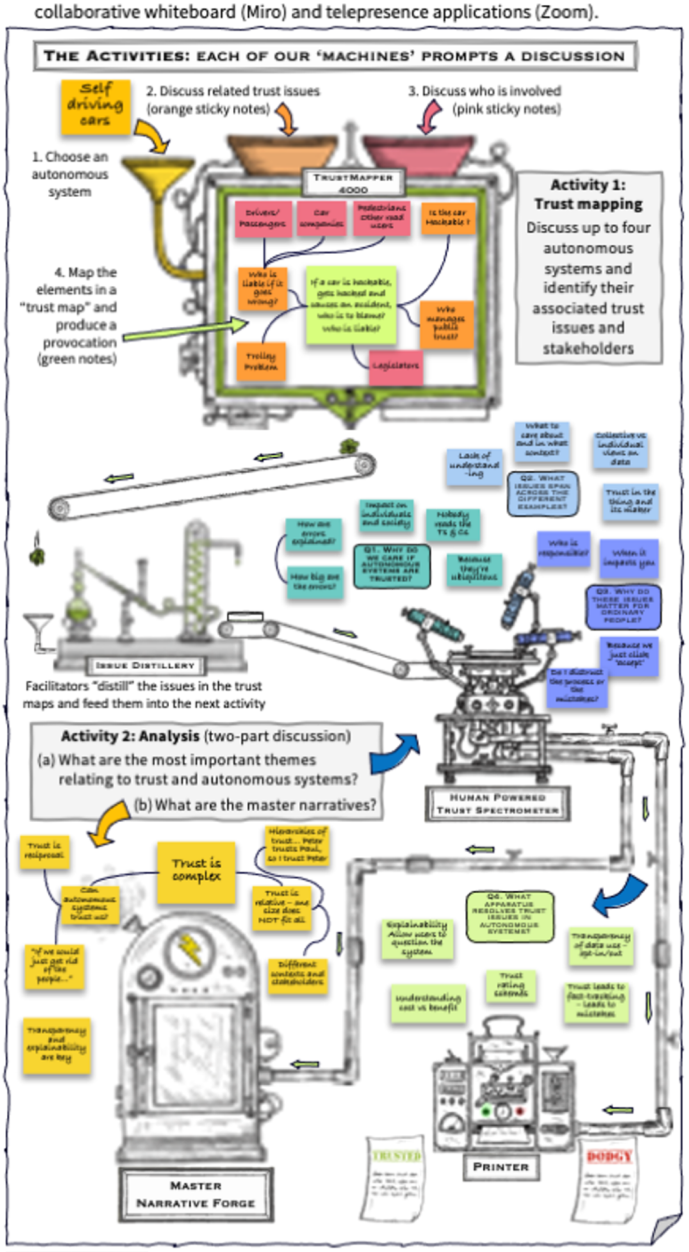
The Workshop

Aims

- To identify key themes or 'master narratives' in TAS
- To make sure we are clear about the shape of these narratives, as a first step towards bridging these understandings with the general public.

Method

- Our workshop coordinated and captured a focused discussion about TAS with domain experts (17 in academia + 5 in industry).
- Activities and discussion were designed and facilitated through online collaborative whiteboard (Miro) and telepresence applications (Zoom).



Capturing the Narratives

What next?

Post Workshop Analysis

- Theme development
- Share report with TAS members

Develop Speculative Designs and Consult with Publics

- Develop speculations which distil, embody, and communicate key positions, ideas and issues
- Deliver these to public groups in collaboration with Ruskin and Design Museums

Reflections and Reporting

- Construct 'master narratives' which negotiate between expert and lay positions
- Share and report to TAS members

Preliminary findings

- Multiple hierarchies, contexts, stakeholders and technologies always interact in any given circumstance
- Trust changes over time
- Just because Alice trusts Bob, and Peter trusts Bob, doesn't mean that Peter should or would trust Alice

Trust is a Distributed Concern

THE RUSKIN

acm

the DESIGN MUSEUM

University of Nottingham UK | CHINA | MALAYSIA

Lancaster University