"Trust equals less death - it's as simple as that" : Developing a Socio-technical Framework for Trustworthy Defence and Security Automated Systems

Asieh Salehi Fathabadi
A.Salehi-Fathabadi@soton.ac.uk
University of Southampton
UK

Pauline Leonard
Pauline.Leonard@soton.ac.uk
University of Southampton
UK

ABSTRACT

With the dramatic rise in the affordances of Automated Systems (AS) across the full range of industrial sectors, designing and implementing AS which are judged as trustworthy by their users is a key challenge facing systems developers, industrial managers and employees alike. However, for some domains, such as Defence and Security (DAS), the stakes are particularly high: a failure of the system could result in fatalities in significant numbers. Gaining a better understanding of the sociological and technical foundations of trustworthiness is critical for the sector, essential for both building trust and designing robust technical solutions to maintain this trust. This paper draws on new interdisciplinary research conducted in the Defence and Security (DAS) sector, exploring social and technical conditions and understandings of trustworthy automated systems. We argue that the distinctiveness of DAS brings some very specific challenges to both developers and users, but the findings of the research have also relevance to a wider variety of domains, especially those where the outcomes may be a matter of life and death.

CCS CONCEPTS

• **Human-centered computing** → *HCI theory, concepts and models.*

KEYWORDS

socio-technical, trustworthy, trust, automated systems, AI

ACM Reference Format:
Asieh Salehi Fathabadi and Pauline Leonard. 2024. "Trust equals less death - it's as simple as that" : Developing a Socio-technical Framework for Trustworthy Defence and Security Automated Systems. In *Second International Symposium on Trustworthy Autonomous Systems (TAS '24), September 16–18, 2024, Austin, TX, USA*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3686038.3686071

TAS '24, September 16–18, 2024, Austin, TX, USA

ACM ISBN 979-8-4007-0989-0/24/09
https://doi.org/10.1145/3686038.3686071

1 INTRODUCTION

Developing trustworthy Defence and Security Systems (DAS) is a key challenge facing contemporary political economies, as international environments face increasing instability. Technology is changing the face of DAS and countries such as the UK are investing billions in their design [14]. In many ways DAS is a 'unique' sector, as the Autonomous Systems (AS) include those designed not only to enhance efficiency and productivity as in other sectors, but also civilian security and peacebuilding, as well as those to support weapon use, and sometimes weapons themselves [16]. Increasingly, national security focuses on military preparedness, including advanced weapon AS [17]. The consequences of automated system use in DAS may therefore include destruction of property and other devastating impacts on human life- and even death. Much of DAS work is time-critical, such that AS may need to be deployed without delay, without the user pausing to think "do I trust this to work?" or wondering "what will I do if it doesn't work?" The social conditions of trust, understanding what may determine the trustworthiness of an automated system, and how this may be broken and then repaired are critical issues of the sector., impacting on the effectiveness of decision-making and even preventing further violence and conflict.

However, 'trust' and 'trustworthiness' are challenging and complex concepts, with multiple definitions and theoretical explanations. For some, the concepts are ultimately incoherent and indefinable, while for others pinning down their essential components is a key scientific task. The concepts have expanded in meaning and application with the development and rapid implementation of new digital technologies, particularly AS where machine decision making is to be trusted to replace human decision-making. The shift towards AS is underpinned by promises of enhanced performance benefits, in terms of speed and accuracy of data analysis and higher levels of safety. However, sociologically, different levels of stakeholder knowledge, understanding, experience and imagination of AS, and different amounts of agency in terms of technical ability and responsibility, lead to different levels of trust towards the deployment of new systems, and therefore, ultimately their use. On the one hand, distrust may result in under-utilisation of the system, while, on the other, high levels of trust might lead to inappropriate or misjudged over-reliance on the system: in other words mistrust. Further, trust is an unstable concept, easily lost but not so easily regained, and rarely 'fixable'. Levels of trust are also highly contingent, differing across time and space, depending on social, political, economic and psychological factors.

The sociological understandings of trust as uncertain and dynamic stand in some tension with the demand for stable, formal verification methods that require a clear-cut specification of 'desirable properties' to ensure compliance with defined specifications, such as demanded within DAS..

This paper adopts a sociotechnical approach to develop a framework of understanding to the social and technical factors involved in building and developing trust in AS within the Defence and Security Sector. While much in-house research is being done on issues of trust and trustworthiness, an academic 'outsider' perspective was welcomed to enhance and support knowledge and understanding. This research drills into the nature of trust in DAS: the elements involved if AS are to be trusted to replace humans for crucial tasks in the future.

It was with this objective that we designed our mixed methods, interdisciplinary research. Working in partnership with a key DAS stakeholder, the research objectives were to investigate:

(1) Stakeholders' understandings and definitions of trust.
(2) Stakeholders' understandings of trust in operational context.
(3) Existing levels of trust in AS.
(4) Levels of understanding of technical solutions that safeguard trustworthiness.
(5) Impact of knowledge of technical solutions on trust levels.
(6) Perceived challenges for trustworthiness of AS and their adoption.

The paper is structured as follows: Section 2 discusses the state of the art and the motivations for our study. Section 3 presents the study specifications. The sociological analysis is provided in Section 4, followed by the technical analysis in Section 5. The discussion, including limitations and recommendations, is presented in Section 6. Finally, Section 7 concludes the paper.

2 RELATED WORK AND MOTIVATION

The concept of trust is complex and multifaceted, non-material and abstract, and subject to multiple definitions and disciplinary understandings. Sociologically, trust is variously conceptualised to foster co-operation [12], reduce social complexity [25] and alleviate uncertainty and vulnerability [26]. Trust is often perceived as involving an almost faith-like quality regarding other people's actions and intentions [37], which involves accepting the risks inherent in a given relationship [35], and a willingness to become vulnerable to another person or actor [33]. More critically, trust can also be seen as an outcome of ignorance or uncertainly with respect to the unknown or unknowable actions of others: we may have little option other than to trust in certain contexts where we have limited knowledge or skills, such as surgery or plumbing or, indeed, automated systems, AI and robotics.

The sociological approach sees trust as an ongoing cooperative process or practice to be sustained between interdependent agents through social action and relations. This understanding competes with more stable and deterministic interpretations of trust as a psychological state, or 'attitude' [15]; or as a behaviour, based on a rational decision to accept vulnerability [3, 24]. In contrast, social constructionist approaches stress the undetermined and contingent nature of trust and trustworthiness: trust and the evaluation that someone or thing is trustworthy 'may rest on particular reasons but is not explained by them' [37]. In a similar vein, social technical studies' (STS) approaches acknowledge the complexities of factors which need to be considered in trust relations between people and technologies such as AS. These include, at the very least, social identities, culture, power, politics, ethical and legal issues, organisational contexts, work roles and career aspirations [7, 34]. The uncertain nature of the human/non-human relationship across these realms means there is no predictable (trust) outcome which will result from any given exchange or activity involving technology [5, 19]. Through interactions in specific spaces and over time, we either build confidence in a person, object or system, leading to higher levels of trust or, if someone/thing does not prove they are worthy of trust, trust dwindles [13]. Trust is thus temporal in nature: an ongoing, dynamic process [27], 'located in the present continuous, something that goes on being made and renewed' [3]. This approach to trust therefore holds that trust in AS is not significantly different from trust in other people: both are unstable and dynamic, and built through ongoing interactions and use.

How can we capture this complexity when it comes to trustworthy autonomous systems in DAS? This paper takes a socio-technical approach to answer this question in three parts.

First, we approach the question sociologically, looking particularly at the first three of our research questions to show that it is useful to identify the different *discourses of trust* as used by the stakeholders themselves, within specific contexts. In this paper we use the term discourse in two ways: first in the more straightforward sense of 'language at text level' [8] - spoken or written language in use. The second, and primary use of discourse here is as a form of social/ideological practice [10]. Discourses are forms of knowledge or powerful sets of assumptions, expectations and explanations, governing mainstream social and cultural practices. They are systematically ways of making sense of the world' [6]. In Section 4 of the paper, we unpick the different discourses used by our respondents, all stakeholders working in different operational contexts to describe trust, and the trustworthiness of the AS they use.

Second, we look more specifically at the fourth and fifth of our research questions, concerning the *technical solutions*. Technical approaches to meet trust and safety concerns are widely investigated yet under-used in the context of measuring users' trust in autonomous AI systems. Interdisciplinary socio-technical approaches, grounded in social science (trust) and computer science (safety), are less considered in AS investigations.

Finally, we bring our findings together to discuss the perceived challenges for the trustworthiness of AS. We conclude by arguing that interdisciplinary socio-technical approaches, grounded in social science (trust) and computer science (safety), offer important and valuable insights to understanding the complexities of trust and trustworthiness.

3 THE STUDY

The paper is drawn on research conducted within two related research projects funded by UKRI's TAS from 2021 to 2024. The broader aims of VESTAS (Verifiably Safe and Trusted Human AI Systems) and HANA-HAIP (Harnessing Trust and Acceptance in Human-AI Partnerships) are to provide a research roadmap that

present the challenges and technical solutions for the design and development of safe and trusted Autonomous Systems. Working collaboratively with our industrial partner, an organization with expertise in defence and security systems, which we give the pseudonym DSAS, our interdisciplinary approach combines social science conceptualisations of trust with computer science approaches to safety. The inclusive stakeholder contributions expand the domains of application and the diversity of perspectives, to inform the development of trust techniques and interventions for subsequent trial, and policy recommendations for regulators. Given the national security issues involved in the research, our access to DSAS was an unusual privilege. We were aware that there would be areas which employees would not be able to discuss, but we were keen to use semi-structured interviews to enable in-depth understanding of everyday work practices and conceptualisations of trust and trustworthiness [30]. Recruiting to the interviews took time and patience, aided by a research director within the organisation. Over a period of a year, we have conducted a total of ten hour-long interviews online, with a sample carefully recruited to meet our diversity requirements of:

- Domains (Air Force, Army, Navy)
- Roles (regulation, software dependability, AV design, flying, training)
- Rank/status (Majors, reservists)
- Experience (20+yrs of military service; front line personnel 12+yrs; reservist for 6yrs)
- Engagement with AS (designers, operators, trainers, "vague understanding" (P2)

The data collection and analysis were conducted iteratively.To date,our respondents include: :

- P1: (m[1]) 23yrs in Air Force, frontline, Chinook; software engineering, test engineer; MOD research community, safety and software dependability; regulator
- P2: (m) Major, Army, 12yrs service. Military advisor, combat role, Special Ops. Afghanistan
- P3: (m) Major, Army Air Corps. Helicopter pilot, instructor, supporting Special Ops
- P4: (f) Army reservist, Air Corps. 6yrs at DSTL. Human factors. Apache helicopters
- P5: (m) 13yrs at DSAS, PhD in autonomous submarine vehicle design
- P6: (m) Human Factors Engineer
- P7: (f) Principal Analyst, Researcher in Human-Autonomy Teaming
- P8: (f) Principal Analyst, Human Autonomy Teaming Project Technical Authority
- P9: (m) Project Technical Authority, Global Force Projection (AW), Principal Advisor, Gun Systems
- P10: (m) Head of Innovation, Counter Terrorism Policing

To answer our research questions, we first coded our data according to our five research questions:

- RQ1: Definitions of trust : how is trust understood generally, in social contexts?
- RQ2: Trust in Autonomous Systems (AS), in an operational context.
- RQ3: Existing levels of trust in AS.
- RQ4: Levels of understanding of technical solutions that safeguard trustworthiness.
- RQ5: Impact of knowledge of technical solutions on trust levels.
- RQ6: Perceived challenges for the trustworthiness of AS.

We now turn to discuss these more fully, turning first in Section 4, to sociological issues of trust in AS and in operational context. We then address technical issues, in Section 5, before summarising the perceived challenges for the trustworthiness of AS in Section 6.

4 SOCIOLOGICAL ANALYSIS

For this section of the analysis, while our participants demonstrated diversity in the range of the discourses they used to discuss these questions, our initial reading of the transcripts suggested some broad systematic patterns and positions in the responses. A vision of a typology started to emerge, inspired by the theoretical literature on trust and enriched as the data set grew [28]. We then coded the data for questions 1-3in greater depth, looking for themes in participants' accounts of how, when and where they would trust in general social contexts, in terms of AS, and in terms of using AS in operational context. This affirmed three overarching discourses of trust were actively constructing and mediating experiences of AS. As self-reflexive researchers, we are aware that these are discursive constructs that we choose to categorise and foreground in this study, rather than arguing that these are universally self-evident in positivist terms [6]. The three discourses can also be conceptualised as relating to each other in the form of 'levels of trust', as follows:

- Level 1: Absolute Trust : complete and unquestioned trust in person or AS
- Level 2: Interactional Trust : conditional, dynamic trust gained through interaction with person or AS
- Level 3: Self-based Trust : trust gained through personal agency and control over person or AS

This initial typology provided a valuable 'conceptual-empirical' tool by which to analyse the respondents' positions and what these meant in practice in the specific context of DSAS. The typology is intended to be illustrative, not exhaustive: an overarching frame to represent the diversity of positions in combination with the theoretical knowledge [28]. As well as being a valuable 'descriptive tool' [4], the typology might also be useful as a thinking tool for DSAS' own strategic planning.

Lessons Learned:

Analysing the data according to the three discursive levels revealed how there were intersections between roles, domain and approaches to trustworthiness, with those located in similar contexts and at similar levels in organisational hierarchies sharing broad positions. While most demonstrated consistency in their positioning, there were occasional shifts in position from the same participant, articulating a mix of the levels of trust. From our interviews across the domains, we found that those in less senior positions tended towards more 'absolute' discourses of trust. These tended to be based in the land-based domain, trained to accept and act on orders from more senior officers without question. Those

[1]f: female, m: male

with more personal agency for their work, trained to make and take decisions according to their own judgements, tended towards discourses of 'self-based' trust. These participants tended to be located in the Air Force, very much 'at the sharp end' such as pilots, who needed to know that ultimately they had control over the AS and could determine how and when it is used and what to do if it fails. In other words, the AS was subsumed to trust in their own personal skills and capabilities. There was also a 'middle ground' , occupied by those in technical positions, security and intelligence who were keen to suspend judgment on how much to trust the AS, preferring to work with and constantly evaluate the AS. We now turn to discuss these findings in more detail.

4.1 RQ1: Definition of trust

Our discussions in the opening sections of the interviews penetrated our participants' understandings of trust and trustworthiness more generally, in social contexts, as they responded to the question: what does trust mean to you? How would you define it? This was the topic of most consensus, with most agreeing that trust could be defined through concepts such as 'reliability' and 'confidence'. Some drew a distinction between 'trust' and 'trustworthiness', with P8 noting:

> "Trust is what a person feels about a person or system, you can say you don't trust something, but you could acknowledge it is trustworthy." (P8)

As we dug further into the interview transcripts, the data revealed three broad positions, which corresponded to the three discursive levels of trust which we identified above. Within the first level, *Absolute Trust*, participants revealed that trust is gained through being able to completely 'hand over' agency and control to another person or thing:

> "Trust is the ability to be hands off." (P5)
>
> Being able to rely on someone to do the task that's been assigned to them and in the allocated timeframe." (P2)
>
> "Your reliance on something. When you expect someone or something to act in a certain way and they do. How you think something or someone will behave." (P4)

Themes of reliability and confidence underpinned this level of trust, where, once established, Absolute Trust is demonstrated by a relinquishing of personal agency. However, for others trust is built through ongoing interaction, always conditional and dynamic. Within this discourse, the second-level discourse of *Interactional Trust*, personal agency hovers in suspension, like a 'hand on the gearstick' , the participant is ready to take it back if trust is lost:

> "There's an initial base level and then it can be either built up or taken away depending on your interaction with a given individual or group."(P3)
>
> "Optimal trust would look like a well-working human-human team. You know you can rely on them. If they fail, it's because I've asked them to do something beyond their ability. You don't lose trust in them: you flex and work alongside those fallible elements." (P1)

The fact that we were talking to people who constantly work in challenging contexts was clear. This is a world where trust is hard-earned, as P2 reveals:

> "Trust can only be engendered through hardship or significant amounts of training and investment of time into personal relationship stuff." (P2)

Some participants were clearly more reluctant to ever hand over control unquestioningly to another person or thing. This third level of *Self-based Trust* was founded on ongoing personal agency, and never quite losing a sense of control:

> "It's me knowing… the system works." (P2)
>
> "You want to be able to check and understand the constraints where they exist and be able to map those through. And that then results in a trustworthy system, one that you can understand and understand where it's failed and where it hasn't done what you might have expected to do, which is inevitable because that's the way systems work". (P10)

P10 notes that there is a powerful underlying assumption that systems 'should be able to be taken at face value', that is, that they can be trusted absolutely, but he feels that this assumption 'can often be quite wrong'. For P10, the most valuable skill is personal understanding and 'informed decision-making' rather than trust, an observation which underscores the importance of the self in decisions about the trustworthiness of AS. Inevitably therefore, discussions of definitions of trust melted into discussions of trust in AS in operational context, to which we now turn.

4.2 RQ2: Trust in AS in operational context

Discussions of trust in AS were clearly central to our participants' everyday working concerns. This is an issue of critical importance to DAS, where people work in extreme contexts where a moment's hesitation may lead to fatalities. Aware that this may mean that trust in AS will be hard-earned and potentially fragile, the sector itself spends significant amounts of time and funds investigating this issue. Many of our participants were involved in activities aimed at understanding and improving the components of trustworthy AS. Nevertheless, differences in positions taken towards the trustworthiness of AS were revealed in the data, with participants located at all three levels of trust.

The hierarchical chain of command for which the DAS sector is renowned underpinned those demonstrating high levels of Absolute Trust.

> "There is trust through a chain of command. I trust something because I'm told to." (P1)

The AS in use in DAS are predominantly high-cost, highly specific tools with high levels of secure data. The consequences of losing these to the wrong hands was bound up with the need for absolute trust not only in the system itself but for its safe return, as P5 reveals:

> "The confidence that the kit will come back, whether or not the mission is completed." (P5)

The critical nature of some of our participants' working lives resonated through their justifications of the need for this form of trust. For them, the bottom line is stark:

> "Trust equals less death. It's as simple as that." (P2)

One of our respondents whose work directly involves researching the complexities of what makes a system trustworthy, noted that people's relationship with other technologies can spillover into AS, perhaps creating Absolute Distrust:

> "There's always someone that has some story about a satnav! It's like, well, I don't trust my satnav. Therefore, I won't trust this!". (P8)

Others were less able to hand over to the AS with such completeness. Continuous interaction and reinforcement are required to prevent a loss of confidence, and as such, the level of Interactional Trust clearly dominated some of the responses. The contingent nature of the AS to the task at hand was clearly revealed in our interview with P4, underpinning the importance of continuous human-system interaction

> "If I tell a drone to turn right, it turns right and it communicates that back to you, gives you that feedback." (P4)

The importance of the testing, the accessibility and explainability of the design, and the confidence built up through multiple previous interactions are the foundations of trustworthiness for those deploying AS in critical contexts:

> "The system's gone through multiple verification and validation stages, and has explainable feedback or transparency, with the option to override it." (P4)

> "If things go wrong, the consequences could literally be fatal. If something's told you something and you've dropped a weapon somewhere and it would, the information wasn't accurate. I think you would find it very hard to trust, you know, if it said, oh, there's a load of bad guys in that bus and it was actually a school bus. You would then find it very hard to rebuild because the consequences are huge (P8).

The sharpness of the context means that others always felt that they needed to be able to rely on themselves first, as well as their fellow teammates, rather than the system. From this perspective, the AS is never more than a tool for human use, rather than an equal or more powerful member of the team:

> "If my fellow commander fails to rendezvous repeatedly, I won't trust him anymore. I'll make contingency plans. If an AV lets me down, it's easier to discard tech as an option." (P2)

Further, DAS attracts people who want to develop, use and rely on their own high-level skills, not just operate a system, and hand over trust to that. Being personally involved in the task is critical to their sense of identity at work, as P8 notes:

> "There are lots of them particularly aircrew, they want to fly the planes, they want to be doing, you know, dropping the bombs or whatever. They don't want all these systems to do it for them, because why then they could be in industry, earn loads of money. (P8)

The dynamic nature of trust revealed at this level underscores the conceptualisation of trust as temporal in nature: an ongoing, dynamic process that goes on being made and renewed [3]. With this in mind, what are the existing levels of trust in AS? It is to this we now turn.

4.3 RQ3: Existing levels of trust in AS

Our interviews revealed significant variance in the levels of trust, with skills in technical knowledge being critical to determining participants' positions. Those locating themselves at the level of Absolute Trust were in the minority, but for some, if a list of criteria was met, then trust would follow, as P1 explains:

> "It's about dependability and the characteristics of system : safe, available, secure, performant" (P1)

> "To trust the capabilities of an AS, it needs to be able to duplicate given behaviours for standardised inputs." (P1)

> "When I'm working with my vehicles, I build up trust when it has repetitive behaviour, so when I tell it to do a certain thing in the same kind of environment and it's worked in before and it does that thing exactly the same, then my trust starts to build up. It's when you get even the slight variations in its behaviour that my trust degrades in the vehicle." (P7)

P7 works with underwater vehicles and while he is highly skilled, he admits that he has a foundation of Absolute Trust in a system which he acknowledges may be misplaced:

> "I think my issue is even over the many years I've worked with them, I still treat them like if they were a dog, so I expect the dog to have some basic intuition, which obviously vehicles don't, and as a software engineer I should know this because I know how they're programmed, but you've always got this… when you put the vehicle in the water there's some kind of instinct that takes over, like you're treating it like some kind of sheepdog, I think. So, yeah, I have to keep tripping myself up." (P7)

Most were quite sceptical of new technology, especially that which is to be used in high-risk scenarios. For a system to be judged as trustworthy requires lengthy and rigorous training, underscoring the importance of continuous interaction and evaluation. As such, levels of Interactional Trust were clearly demonstrated:

> "The tech takes a long time to understand how to operate it. It's hard-won experience. Then you have to go out into the field and do big training exercises to really understand its limitations" (P2)

While it was appreciated that any new technology must go through early iterations to become reliable: "Any new tech has lots of bugs and errors." (P3), the uncertainty as to whether a system would fail suddenly, either through degrading sharply or gradually declining in performance, was ever present. The fear of sub-optimal performance or, at worst, complete failure, was pervasive. Many realised they might have to cope with unexpected outputs, which would mean relying on their own skills. As such, levels of *Self-based Trust* were high:

> "Do I fully understand what it's capable of, do I know where those gaps in my understanding are? How is

> failure communicated? How does it degrade? Is it a sort of gradual decline?" (P3)

The reliance on the self, and one's own skills, means that if these are lacking, the resulting anxiety and confusion will degrade trust in the AS:

> "I lack trust in AS when something unexpected happens, and I can't work out why. It just degrades trust. I don't know if that's going to happen again. And I don't know what caused it in the first place." (P3)

Our participants demonstrated that they tolerate a degree of iteration with the development of trustworthy AS and most are very keen to become subject matter experts. However, what might be tolerated with software is not fully transferred to AS in DAS, particularly for senior officers, who take responsibility for the consequences of failure. Possessing knowledge of technical solutions is often felt to be essential, and it is to these we now turn.

5 TECHNICAL ANALYSIS

Aiming to ensure trust without a clear understanding of the technical aspects may result in infeasible expectations from stakeholders. While much of the research in the AS domain focuses on specific technical problems to achieve safety [1], verifiability is crucial for trustworthy AS due to challenges that arise in the dynamic and uncertain human-AS relationships [20]. We consider verification to be the process of obtaining evidence that a system of interest meets a specified property or properties. In this section, we turn to explore the technical challenges that our participants identified, and the properties needed to solve these challenges.

Our discussions within the 'technical' section of the interviews supported the integral interrelation of social and technical factors , and their respective roles in shaping the conditions conducive to trustworthy system performance. As we noted above, we learned that earning stakeholders' trust in AS necessitates adopting a socio-technical perspective: simply verifying the technical reliability of a system is inadequate for ensuring trust in the AS.

5.1 RQ4: Levels of understanding of technical solutions that safeguard trustworthiness & RQ5: Impact of knowledge of technical solutions on trust levels : Unravelling complexity by explainability

The technical discussion proved the necessity of implementing technical solutions to enhance stakeholders' understanding of the dynamics between safety and trust. . Our research highlighted how different levels of technical background and understanding, can impact trust in AS, and how important it is for our participants that they can not only understand the AS but 'step in' if necessary:

> "Someone with technical background, they would understand what and why the system is doing and then they'll know when to intervene and when not to intervene." (P6)

> "In operational context, do I fully understand what the system capable of? If I don't, do I know where those gaps in my understanding are?" (P3)

> "Providing many different scenarios to try and understand how the systems operate and how they work, what their likely behaviours are in response to different situations." (P3)

Insights from interviews reflect the complex interplay between technical solutions and trust in military operations, highlighting the need for adequate training, resources, and understanding to build reliable AS. In terms of the levels of trust, therefore, these quotes demonstrate Self-based Trust- the importance of being able to rely on one's own skills if things go wrong. Two aspects emerged as key here: Complexity in Autonomy and Explainability.

Complexity in Autonomy: Alongside Self-based Trust, our interviews also revealed how *Interactional Trust* is of critical importance to technical solutions. The technical discussion revealed that trust in AS is influenced by various types of complexity:

> "There are different kinds of complexity that can influence trust: complexity in the human interaction with the system, complexity in the environment where the system operates, and complexity in the tasks the system is designed to perform." (P9)

The three key types of complexity identified by P9 were demonstrated elsewhere in our interviews. They can be elaborated as follows:

- Human Interaction Complexity: How the system interacts with humans can affect trust. If the interactions are too complex or not well understood, trust may be diminished.
- Environmental Complexity: The environment in which the AS operates can impact trust. If a system performs well in a controlled environment but fails in a more challenging one, users may lose trust.
- Task Complexity: The complexity of the tasks the system is designed to perform can influence trust. Simple tasks may inspire more trust compared to complex ones where the system's decision-making is less transparent.

The complexity of interactions, environment, and tasks, can significantly influence trust in AS. To address this, our data shows that understanding and mitigating these complexities through better explanations can enhance user confidence and trust.

Explainability: Our interviews also revealed that trust can be enhanced through a better understanding of the technical solutions behind AS. This understanding can be developed from:

- Confidence Through Knowledge: : detailed explanations of how the system works and knowledge of mathematical verification can boost user confidence.

  > "The more you know how a system works, the more confidence and trust you'll have in it and into yourself to to utilise it." (P9)

  > "Trust would be integrated with good situation awareness of what the system (algorithms) is doing, and why it's doing it?" (P6)

  > "A lot of feedback is about explainability and the transparency of a system." (P8)

- Experience and Exposure: Practical examples highlighted that training, play a crucial role in building trust. Users need

to understand not only how a system works but also its limitations:

> "We put automation into the system to help the user, but then we didn't give them sufficient training." (P9)

When providing a technical explanation and training, our findings show it is important to consider the audience's understanding. Our respondents argued that the main focus should be on clarifying what the system aims to achieve and why it might have fallen short or encountered limitations in its process. It is necessary to break down the system's goals and constraints in a way that is easy for them to grasp:

> "You would want an explanation the users can understand. And I think the key thing is explaining in their language, kind of what the system is trying to do and how it may have failed or what the constraints are in the way that it's processed it." (P10)

Lessons Learned:

From our analysis of the data in our study, we would propose to use Public Engagement (PE) techniques as a potential solution to bridge the gap between complex AS system behaviours and end users' understanding. By engaging the DAS community and involving them in the AI development and deployment processes, PE techniques can elucidate these behaviours and enhance user trust and transparency [32]. These techniques will include:

- Enhancing Transparency through Public Engagement: Public engagement involves a spectrum of activities from informing the team about decisions and policies to actively involving them in decision-making processes. By employing PE techniques, we can demystify complex AI behaviours for end users, fostering greater transparency and trust. This participatory approach helps in building a community that is well-informed and engaged with AS technologies [2].
- Effectiveness of Visualisation Techniques: Utilising visualisation techniques in public engagement is a powerful tool to make complex AS behaviours more comprehensible. Visualisations can transform intricate data into accessible formats, making it easier for end users with varying technical backgrounds to understand and interact with the system. This can significantly enhance trust and confidence in AS systems as users feel more informed and empowered [22].
- Addressing the Needs of Non-Expert Stakeholders: Our findings highlight the disparity between the desired properties of non-expert stakeholders and their capacity to articulate these needs precisely in a technical specification language. For those members of the team unable or unwilling to develop expert skills, effective communication and explainability are crucial in bridging this gap. PE techniques, especially those involving visual tools and synthetic imagery, can play a pivotal role in making complex data accessible and understandable to non-experts [21].
- Building Trust through Meaningful Indicators: As one of our interviewees emphasised, using visual tools to present more meaningful indicators to operators can help in building and maintaining trust in AS. Operators often struggle with understanding probabilities, graphs, and numerical values. Providing them with intuitive visualisations helps them to grasp the system's behaviours better, thereby enhancing their trust and confidence in the system's reliability [9].

  > "What we're looking to build upon is more meaningful indicators to operators through visualizations." (P1)

  > "Now this is the challenge in that a lot of operators don't understand probabilities. They don't understand graphs. They don't understand kind of numerical values. And so what we have seen and what we're looking to build upon is more meaningful indicators to operators through visualisations." (P1)
- Fostering Informed Participation: PE techniques that include visualisations understandable to different stakeholders can foster informed participation across diverse domains. By presenting AS behaviours in a visually engaging and easily digestible manner, we can ensure that all stakeholders, regardless of their technical background, can contribute meaningfully to the decision-making process [11].

Overall, our lessons learned underscore the importance of using public engagement and visualization techniques to enhance transparency, trust, and effective communication in AS systems. These approaches can bridge the gap between complex technical solutions and the diverse expectations of stakeholders, leading to more robust and user-centered AI systems.

5.2 RQ6: Perceived challenges for the trustworthiness of AS : Exploring non-linearity, unpredictability, and instability

Our findings have shown how human characteristics of technical knowledge and experience, made up from a combination of familiarity, technical background, and generational differences, significantly impact trust in AI systems. Consistent with the arguments of this paper, trust in these systems is a complex and multifaceted issue, influenced by a variety of factors that often extend beyond the criteria established by designers and engineers. Determining whether a system is trusted depends on how stakeholders make everyday judgements about trust. These judgements may align with, but also notably diverge from, the technical assessments of trustworthiness. Trust in technical solutions is inherently context-dependent, varying according to specific circumstances and occasions.

In DAS, stakeholders frequently utilise AS, and their levels of trust, and assessments of trustworthiness are deeply rooted in their personal experiences. These interactions with the technology can lead stakeholders to question its reliability. However, even when acknowledging its shortcomings, they may still find the system sufficiently trustworthy to consider it useful. This dynamic highlights the non-linearity, unpredictability, and instability inherent in trust within the realm of AS.

The need for users to shift from expecting predictable system behaviour to trusting that the system will achieve desired outcomes, acknowledging that AS and AS might not always behave as anticipated:

> "What we need to do is we need to get users away from the concept of being able to predict exactly how something will behave all the time, into trusting its

> outcome will be achieved and where that has not been achieved, understanding why it's not been achieved." (P1)

The following quote expresses a preference for simpler systems that consistently work well over more complex systems that may be prone to slowdowns or inefficiencies at critical moments:

> "Part of me wants a simple system that works better, more than a complicated system that was less effective." (P6)

There is also a recognition of the unpredictability in user behaviour, particularly in high-stress or critical situations where human judgement may override algorithmic recommendations:

> "When you're really tired and you're overloaded. It's just quite nice to go with it, but then also there are certain grey areas where you're like, oh, maybe it's not perfect." (P6)

> "The technical understanding of soldiers regarding new AS is generally minimal. They can operate the systems but often don't understand the underlying mechanisms, which can affect trust in these systems." (P2)

> "Unlike human failures, technological failures can be harder to understand and predict, making it easier to discard technology than to discard a human for repeated failures. This unpredictability makes soldiers more cautious about relying solely on AS without thorough testing and integration into their routines." (P2)

The understanding and interaction with the system evolve over time and with exposure, and this influences how trust is built, maintained, or lost.

If a verification result is not promptly provided, a stakeholder may lose trust in a system, highlighting the necessity for an efficient trust verification procedure.

Ensuring safety at the present moment does not guarantee sustained trust from stakeholders in all future moments. This underscores the potential relevance of employing temporal logics [32] and sequential decision-making setting to analyse how system assertions evolve over time.

Lessons Learned:

Throughout our exploration of non-linearity, unpredictability, and instability in trust within Autonomous Systems (AS), we have gathered several key insights. These lessons highlight the importance of understanding and addressing the human factors that influence trust in AS systems. By integrating these lessons into the design and development processes, we can create more reliable and user-centered AS.

- Recognizing Human Behavior Variability: In contrast to the linear, structured logic, or syntactic precision typically associated with formal verification methods, we discovered that human behaviours within AI systems can exhibit non-linearity, unpredictability, and inherent instability. These behaviours often do not follow predictable patterns or adhere to straightforward logical frameworks. Instead, they can vary widely based on individual experiences, contextual factors, and situational variables, making them difficult to anticipate and model accurately [38].
- Implications for Trust Preconditions: This unpredictability and non-linearity in human behaviours have significant implications for the development and trust preconditions of AS systems. During system development phases, establishing trust preconditions typically involves defining clear, objective criteria that the system must meet. However, the inherent instability in human behaviours means that these criteria might need to be flexible and adaptive to accommodate the diverse ways users interact with and perceive the system [18].
- Performance in Real-World Settings: A system designed with a specific set of trust metrics might perform well in controlled environments but fail to inspire the same level of trust in real-world settings where human behaviors and interactions are more complex and less predictable [23]. This necessitates a more holistic approach to system design and verification, one that considers not just technical performance but also the nuanced, often subjective experiences of end-users [31].
- Holistic Approach to Design and Verification: By acknowledging and addressing the non-linearity and unpredictability of human behaviours, developers can create AS systems that are more robust, adaptable, and ultimately, more trusted by their users. This approach involves understanding the psychological and contextual factors that influence trust and designing systems that can accommodate these variations [36].
- Adaptive Trust Criteria: Given the variability in human behaviour, trust criteria need to be adaptive and context-sensitive. This means that systems must be designed to learn from user interactions and adjust their behaviour accordingly to maintain and enhance trust over time [18].
- User-Centred Design: The importance of a user-centred design approach is paramount. This involves actively engaging users throughout the development process to understand their needs, expectations, and experiences. Such engagement ensures that the system is not only technically sound but also aligns with the users' mental models and trust expectations [29].

6 DISCUSSION

The findings from our technical solution discussions (Section 5) predominantly align with *"Level 2: Interactional Trust"* (Section 4). This level of trust is characterised by its conditional and dynamic nature, established through ongoing interactions with a person or an AS. Interactional trust evolves as users gain experience and familiarity with the system's behaviour in various contexts. Our results indicate that trust in AS is significantly influenced by the quality and consistency of these interactions, underscoring the importance of designing systems that are responsive and reliable in real-time user engagement scenarios. This dynamic trust is crucial for ensuring that users feel confident in relying on AS for critical tasks [18].

Implications of Explainability on Interactional Trust Explainability plays a pivotal role in fostering interactional trust. When an AI system can clearly articulate its processes and decisions, users are

more likely to understand and trust its behaviour. This transparency reduces uncertainty and scepticism, allowing users to interact with the system more confidently. By demystifying the system's inner workings, explainability helps users form a mental model of how the system operates, making interactions smoother and more predictable. Consequently, explainability enhances the user's ability to trust the AI system conditionally and dynamically as they become more familiar with it.

Implications of Non-linearity, Unpredictability, and Instability on Interactional Trust The non-linearity, unpredictability, and instability of AS systems can significantly impact interactional trust. Users may find it challenging to trust a system that exhibits erratic or unexpected behaviour. To mitigate this, AS systems must be designed to manage and communicate these complex behaviours effectively. Implementing robust feedback mechanisms that inform users about the reasons behind certain actions or unexpected outcomes can help in maintaining trust. Additionally, ensuring consistent performance across various conditions can reduce the perceived instability of the system. By addressing these aspects, developers can create AS systems that maintain interactional trust through reliable and predictable interactions, even in the face of inherent complexities.

Limitations:

Despite the real-world context of our study in DAS, there are limitations that must be acknowledged. The specific nature of these environments may not fully capture the diverse range of user experiences in other domains. Additionally, the high-stakes context of DAS might introduce unique stressors and expectations that could influence trust differently compared to more benign settings [32].

Recommendations:

Based on our findings, we recommend the following to enhance interactional trust in AS systems:

- Prioritise Explainability: Developers should integrate explainability features that allow users to understand the decision-making processes of AS systems. Clear and accessible explanations can build user confidence and trust [21].
- Enhance Feedback Mechanisms: Implement feedback systems that provide users with real-time information about the AI's actions and reasoning. This can help mitigate the unpredictability and perceived instability of AS behavior citeParasuraman1997.
- Conduct User-Centered Design: Engage users throughout the design and development process to ensure the AS system meets their needs and expectations. Understanding user perspectives can help tailor the system to build and maintain trust [36].
- Test in Diverse Scenarios: Validate AS systems in a variety of environments beyond the DAS context to ensure they perform reliably and consistently. This can help identify and address any issues that may arise in practical applications across different domains [11].

By adopting these recommendations, developers can create more robust, trustworthy AI systems that foster strong interactional trust with users.

7 CONCLUSION

This paper presents findings from a study investigating trust and trustworthiness of AS in the Defence and Security Sector. Our analysis of our data to answer our research questions provides significant and original contributions to knowledge of levels of trust in AS in the distinctive, high-risk context, where dis/trust in the AS can lead to critical consequences. The paper draws on an innovative social-technical approach to offer both conceptual and practical contributions and recommendations.

Conceptually, our identification of three discourses or levels of trust: Level 1 Absolute Trust: complete and unquestioned trust in person or AS; Level 2: Interactional Trust : conditional, dynamic trust gained through interaction with person or AS ; Level 3: Self-based Trust : trust gained through personal agency and control over person or AS; helps to provide a framework by which to deconstruct the complexities involved in trust-building. This proved to be productive when drilling in to stakeholders' understandings of trust in operational context where the critical importance of context was revealed. Where a person is located in terms of domain and organisational structure feeds into levels of trust. For example, a team member of the armed forces who is trained to take orders without question is more likely to demonstrate Absolute Trust in AS if instructed to do so. A soloist flying an airplane in a warzone needs to know that they can rely on their own skills and judgement if the AS fails: for them, Self-based Trust is paramount. Investigating RQ3, existing levels of trust in AS, revealed how Interactional Trust pervades across domains and structures: participants working in critical, unpredictable contexts are constantly evaluating, critically assessing and making time-constrained judgements on the trustworthiness of the AS. As such trust and distrust hang in a fine and precarious balance. This was further demonstrated through our findings to RQ4 which looked at levels of understanding of technical solutions that safeguard trustworthiness. Our research shows how different levels of technical background and understanding can significantly impact trust in AS, and how important it is for our participants that they can not only understand the AS but also have the skills and expertise to intersect if necessary. Self-based Trust is therefore also important here. In terms of RQ5, the impact of knowledge of technical solutions on trust levels. our research shows how social characteristics of technical knowledge and experience, a complex combination of familiarity, technical background, and generational differences, significantly impact trust in AS systems.

Our paper demonstrates that trust in autonomous systems in DAS is a complex and multifaceted issue, influenced by a variety of factors that routinely extend beyond the criteria established by designers and engineers. While this presents challenges for trustworthiness of AS and their adoption, our socio-technical analysis helps to provide some guidelines and recommendations for building trust in this context. It is clear that striving for the level of Absolute Trust is not only unworkable but unviable. People in the DAS sector are, in the main, highly trained and highly skilled employees in risky and dangerous environments. In such sharpened contexts, they want and need to deploy their own assessments and rely on their own skills. Working with these levels of Interactional Trust and Self-based Trust is therefore an essential requirement for establishing effective and trustworthy AS in DAS.

ACKNOWLEDGMENTS

This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) through the UKRI Trustworthy Autonomous Systems Hub (EP/V00784X/1).

REFERENCES

[1] Michael Akintunde, Victoria Young, Vahid Yazdanpanah, Asieh Salehi Fathabadi, Pauline Leonard, Michael J. Butler, and Luc Moreau. 2023. Verifiably Safe and Trusted Human-AI Systems: A Socio-technical Perspective. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems, TAS 2023, Edinburgh, United Kingdom, July 11-12, 2023*. ACM, 56:1–56:6.

[2] S. R. Arnstein. 1969. A Ladder of Citizen Participation. *Journal of the American Institute of Planners* 35, 4 (1969), 216–224.

[3] Rachel Ayrton. 2020. The case for creative, visual and multimodal methods in operationalising concepts in research design: An examination of storyboarding trust stories. *The Sociological Review* 68, 6 (2020), 1229–1249.

[4] Kenneth D. Bailey. 1994. *Typologies and Taxonomies: An Introduction to Classification Techniques*. Sage Publications, Thousand Oaks, CA.

[5] Chris Baldry. 2011. 'Editorial: chronicling the information revolution'. *New Technology, Work and Employment* 26, 3 (2011), 175–182.

[6] Judith Baxter. 2003. *Positioning Gender in Discourse: A Feminist Methodology*. Palgrave, Basingstoke.

[7] Robert Boyd and Richard Holton. 2018. Technology, innovation, employment and power: Does robotics and artificial intelligence really mean social transformation? *Journal of Sociology* 54, 3 (2018), 331–345.

[8] Deborah Cameron. 2001. *Working with Spoken Discourse*. Sage, London.

[9] M. R. Endsley. 1995. Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors* 37, 1 (1995), 32–64.

[10] Norman Fairclough. 2010. *Critical Discourse Analysis: The Critical Study of Language*. Routledge, London.

[11] F. Fischer. 2000. *Citizens, Experts, and the Environment: The Politics of Local Knowledge*. Duke University Press.

[12] Diego Gambetta (Ed.). 1988. *Trust: Making and Breaking Cooperative Relations*. Basil Blackwell, Oxford.

[13] Anthony Giddens. 1994. Risk, Trust, Reflexivity. In *Reflexive Modernization*, Ulrich Beck, Anthony Giddens, and Scott Lash (Eds.). Polity Press, Cambridge, 184–197.

[14] GOV.UK. 2024. PM announces 'turning point' in European security as UK set to increase defence spending to 2.5% by 2030. https://www.gov.uk Accessed: 2024-06.

[15] Russell Hardin. 2006. *Trust*. Polity Press, Cambridge.

[16] Marion Hersh. 2022. Professional ethics and social responsibility: military work and peacebuilding. *AI and Society* 37 (2022), 1545–1561.

[17] Daniel Hoadley and Nathan Lucas. 2018. *Artificial Intelligence and National Security*. Technical Report R45178. Congressional Research Service. https://crsreports.congress.gov/product/pdf/R/R45178.

[18] R. R. Hoffman, M. Johnson, J. M. Bradshaw, and A. Underbrink. 2013. Trust in Automation. *IEEE Intelligent Systems* 28, 1 (2013), 84–88.

[19] Debra Howcraft and Phil Taylor. 2014. 'Plus ca change, plus la meme chose?'-researching and theorizing the 'new' new technologies. *New Technology, Work and Employment* 29, 1 (2014), 1–8.

[20] Nicholas R. Jennings, Luc Moreau, David Nicholson, Sarvapali D. Ramchurn, Stephen J. Roberts, Tom Rodden, and Alex Rogers. 2014. Human-agent collectives. *Commun. ACM* 57, 12 (2014), 80–88.

[21] B. B. Johnson and P. Slovic. 1995. Presenting Uncertainty in Health Risk Assessment: Initial Studies of Its Effects on Risk Perception and Trust. *Risk Analysis* 15, 4 (1995), 485–494.

[22] R. Kosara and J. Mackinlay. 2013. Storytelling: The Next Step for Visualization. *Computer Graphics and Applications, IEEE* 33, 1 (2013), 44–50.

[23] J. D. Lee and K. A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (2004), 50–80.

[24] Peter Ping Li. 2007. Towards an interdisciplinary conceptualization of trust: A typological approach. *Management and Organization Review* 3, 3 (2007), 421–445.

[25] Niklas Luhmann. 1979. *Trust and Power: Two Works*. John Wiley, Chichester. Translation of German originals Vertrauen [1968] and Macht [1975].

[26] Aneil K. Mishra. 1996. Organizational Responses to Crisis: The Centrality of Trust. In *Trust in Organizations: Frontiers of Theory and Research*, Roderick Kramer and Tom Tyler (Eds.). Sage, Thousand Oaks, CA, 261–287.

[27] Guido Möllering. 2001. The nature of trust: From Georg Simmel to a theory of expectation, interpretation and suspension. *Sociology* 35, 2 (2001), 403–420.

[28] Melanie Nind and Sarah Lewthwaite. 2020. A Conceptual-Empirical Typology of Social Science Research Methods Pedagogy. *Research Papers in Education* 35 (2020), 467–487. https://doi.org/10.1080/02671522.2019.1601756

[29] D. A. Norman. 2013. *The Design of Everyday Things: Revised and Expanded Edition*. Basic Books.

[30] Karen O'Reilly. 2012. *Ethnographic Methods* (2nd ed.). Routledge, London. https://www.routledge.com/Ethnographic-Methods/OReilly/p/book/9780415561815

[31] R. Parasuraman and V. Riley. 1997. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors* 39, 2 (1997), 230–253.

[32] G. Rowe and L. J. Frewer. 2005. A Typology of Public Engagement Mechanisms. *Science, Technology, & Human Values* 30, 2 (2005), 251–290.

[33] Oliver Schilke, Martin Reimann, and Karen S. Cook. 2021. Trust in social relations. *Annual Review of Sociology* 47 (2021), 239–259.

[34] Cynthia Selin. 2008. Sociology of the Future. *Sociology Compass* 2, 6 (2008), 1878–1895.

[35] B. H. Shepard and D. M. Sherman. 1998. The Grammars of Trust and General Implications. *Academy of Management Review* 23 (1998), 422–438.

[36] K. Siau and W. Wang. 2018. Building Trust in Artificial Intelligence, Machine Learning, and Robotics. *Cutter Business Technology Journal* 31, 2 (2018), 47–53.

[37] Georg Simmel. 1990. *The Philosophy of Money* (2nd ed.). Routledge, London. Original work published 1900.

[38] D. D. Woods and E. Hollnagel. 2006. *Joint Cognitive Systems: Patterns in Cognitive Systems Engineering*. CRC Press.