A Relational Approach to Trust-Building

Dr Rupert Barrett-Taylor
Dr Maire Byrne
The Alan Turing Institute
96 Euston Road, London NW1 2DB
UNITED KINGDOM

rbarretttaylor@turing.ac.uk

ABSTRACT

The digital sophistication of defence and military organisations is growing rapidly, and they are adopting Artificial Intelligence (AI) for multiple applications. AI is not a unitary actor, but a socio-technical object dependent on relations between components, which will arguably require stakeholders from commanders to logisticians to reconsider what it means to trust relating to an opaque assemblage of technologies. How trust is built between humans and technologies is a contested field, and few studies have used Science and Technology Studies (STS) frameworks to research trust from a structural, sociotechnical perspective. Drawing on the work of Latour and Deleuze allows tacit assumptions about human-machine relations to be unpacked. A relational analysis allows both obvious and underlying structural issues to be analysed in a common framework around assemblages of humans, technologies, and their socio-cultural contexts. In turn this allows trust to be located by those using technology against a specific representation of technology that is frequently understood as a black box.

1.0 INTRODUCTION

The evidence of the first two decades of the 21st century indicates that human participation in war will become more rather than less mediated through information technology. The sophistication of military technology has increased commensurate with the increasing number of decisions it is expected to make, and the growing volumes of data it is expected to process. Defence research is being asked to explain the implications of this technical sophistication including integration and exploitation of AI across the range of military activities. In the future, soldiers, commanders and support staff will be expected to trust the output of a range of AI adjacent technologies in a growing range of situations and tasks, from logistics to combat. Defence researchers are already being asked to explain the nature of a trusting relationship between humans and these technologies, despite AI itself lacking a complete definition. It is a label covering a domain encompassing several technologies (Hagendorff 2020:111) some of which already support defence operations. As noted by Suchman, the fixed label Artificial Intelligence belies a fluidity in the field which obscures the risks and practices posed by individual technologies (Suchman 2023:3).

As discussed throughout this paper, the field of AI suffers from a definitional crisis. Sutrop notes that much of the literature around trust in AI pays scant attention to the differences between AI as a machine that fulfils limited human functions and that which possesses decision-making competency (Sutrop 2019:511). Suchman elaborates on this theme by noting that addressing AI as a unitary actor may be a reason why it holds so much potentially unwarranted power and agency of human-machine interrelations (Suchman 2023:4). Hagendorff notes how this problem produces a gap between technicists and ethicists because it effectively leaves them referring to different objects in their respective analyses. Sutrop also notes there is an open question about whether trust should always be sought in human-AI interactions by differentiating between *trust* and *reliance*. This latter concept is potentially critical to further research. She refers to Bryson, who argues that no human can or should need to trust the current generations of AI because it is not a relationship between true peers. Rather, given the current state of technology the trust relationship should be better framed as between trustors and those institutions that developed the technology rather than the technology itself (Sutrop 2019:511–12).

This paper explores the uncertain domain of AI and what it means to trust in the context of complex and varied underlying technology, as a foundation for wider exploration of the subject. It explores literature addressing trust and the ethics of human-technology relations rather than restricting itself to specifically human-AI relations. Based on the insight of Suchman and Hagendorff above, this paper assumes that AI is a 'black box' which comprises a field of well-established technologies arguably only different in their collective emergent effects which must be unboxed and demystified to be properly understood. Based on this assumption the first sections of the paper explore different ways of conceptualising trust and contrasting frameworks for understanding the human relationship with technology. This provides a framework within which the literature on trust itself can be interrogated. The paper concludes that existing analyses of trust lack a clear theoretical framework to define the relationship between humans and technology whilst descriptions of AI itself lack the specificity which would locate exactly 'what' necessitates a trust relationship.

2.0 WHAT IS TRUST?

2.1 Conceptualising Trust

Trust is a concept subject to considerable variation and interpretation. It can be understood as a proxy for reliability or a more subjective relational problem which requires interpretation and discussion. Users of technology are free to trust but also to distrust and mistrust it. The act of building trust between humans and in this case, AI applications requires a framework to understand the decisions made by users. McKnight calls trust a vital relationship concept central to interpersonal and commercial relationships, and is important where risk, uncertainty or interdependence exist. However, he contends that it has been defined in so many ways across multiple disciplines to suit empirical research that a typology is the best way to understand it as a single definition is impossible (McKnight and Chervany n.d.:827). The interdisciplinary typology suggested by McKnight includes the disposition to trust, institution-based trust, trusting beliefs and trusting intentions (McKnight and Chervany n.d.:829). Even when referred to briefly, this outline of a typology illustrates the complexity of the domain of trust, and the variety of different parties involved. Trust is inherently a problem of ethics, but Reinhardt notes that by itself "… AI ethics overloads the notion of trust and trustworthiness and turns it into an umbrella term for an inconclusive list of things deemed 'good'" (Reinhardt 2023:735–36).

It is clear, therefore that trust, like AI, is both consequential but also ambiguous. It is important but vague, and without specificity, it is difficult to mobilise as a working concept with organisational utility. However, as will be discussed later in this paper, overcoming the ambiguity of trust relationships has frequently been attempted by instrumentalising the trust relationship between the user and material technology. Although a reductive analysis allows for more definition and specificity it also removes the agency of the user in their relationship with technology and makes the trust relationship appear static rather than transient, complex and constantly in becoming.

2.2 Relating Trust to Technology

Baier defines trust as "…reliance on others' competence and willingness to look after, rather than harm, things one cares about which are entrusted to their care" (Baier 1986:259). However, Sutrop observes that the work of Baier differentiates between *trust* and *reliance*. Trust can therefore be thought of more specifically as an interpersonal relationship between peers, whilst *reliance* can be understood as framing relationships with inanimate objects. Trust is conditioned by the potential of *betrayal*, whilst objects will *disappoint*. This statement seems counterintuitive, because some objects can do more than disappoint, for example a parachute that doesn't open is lethal and its failure could be construed as a betrayal. However, as

Grint and Woolgar note, the effect of technology exists only as a result of its fusion with specific human actors (Grint and Woolgar 1992:374). Thus, the user of a parachute can rely on it, but on the basis of trust that those that packed it, maintain it and manufacture it do their jobs properly. Reliance is part of trust, but trust reflects the social network around the material object.

Trust is also distinct from trustworthiness, which is a characteristic of the trust object. In the case of AI, trustworthiness could be thought of as a characteristic of both the black box of AI and of its individual components. Only those who are trustworthy have the power to betray (Sutrop 2019:505), which in the case of a human-to-human relationship can be interpreted straightforwardly. However, if one separates *trust* from *reliance* as suggested by Baier, the concept of trustworthiness becomes necessarily more complex. Reliance is a measure of the relationship between a human and the material character of technology, whilst trustworthiness implies the machine must also somehow symbolise the vulnerability of human relations to result in betrayal. Thus, explicitly or implicitly, trust is a *socio-technical* rather than just a *technical* problem. The reliance of the user on the object of technology is *part* of the network of relations around it which forms the *trust* relationship which is located amongst the institutional links that define the capacity of the technology. This is consequential to understanding how users trust AI applications.

As already noted, AI suffers from a lack of definitional specificity, and its technology is obscured within a black box from which capability emerges. This lack of specificity has already been observed to exacerbate problems interpreting what user trust means in relation to technology. The trust relationship between AI and its human users is arguably not a straight line between the characteristics of the black box and the human. Although the black box of AI may anthropomorphise its capability, it is still not a human and cannot replicate human interrelations, and thus must be treated differently. Opening the black box to reveal a more specific understanding of what AI is constructed from results in more complexity without the anthropomorphising capacity of the box. However, going a step further and adding a layer of social relations to the unpacked box offers the possibility of identifying where in the AI system human-to-human trust relations are. Whilst Suchman's non-specific AI black box is assumed to be a material technological object, unpacking it as a sociotechnical system offers the possibility of understanding more specifically where the human user locates their *reliance* and *trust* in the system.

However, Reinhardt argues that if AI is fully understood through transparency, then trust is unnecessary: "increasing transparency … actually decreases the need for trust by decreasing uncertainty" (Reinhardt 2023:738). Numerous authors advocate for the role of transparency in promoting trust. However, in a situation where a system has been understood in its sociotechnical complexity, it is possible that trust can be located amongst the web of relations between humans and the underlying technology of AI. If this is the case, then the need for transparency is linked to the possibility of trust rather than eliminating the need for it or mitigating for its absence. This paper will explore a framework for mobilising sociotechnical methods to build trust relations by rendering systems visible and specific if not completely transparent.

3.0 SOCIOTECHNICAL SYSTEMS AND TRUST

3.1 Trust Frameworks and Epistemologies

Many of the meta-analyses of trust building between humans and technologies including AI do not interrogate the epistemological bases of their analysis. This limits the scope of subsequent guidance to the individual human-machine relationship rather than considering the wider structural issues which might be also affecting this relationship, not least because the technology is treated as a unitary actor. The analysis falls back on studying the emergent properties of the machine rather than attempting to break this box open in the pursuit of specificity and greater understanding. This criticism is not intended to suggest that the field of trust building with AI and technology is not complex and empirically rich, but it is set in an epistemological framework of positivism. This is true of most of the science and technology sector of the economy and the defence domain which is aligned to it. The defence community has if anything been a historic spearhead for technology investment and adoption and has had a formative role in the values and culture around technology itself. From hardware to software, technology is understood as *determining* change in defence as understood through its essential material characteristics. Through this lens information technology including AI drives improvements to the *efficiency* and performance of defence and military capability through *optimisation* of processes using data and algorithms to process measurable information about the world. Reliance on *technological determinism* and *positivism* as a framework to understand technology comes with analytical limitations which affect how trust can in turn be understood. How this epistemological framework manifests is highlighted in literatures concerned with Managerial Ideology which is characterised by a belief in the universal applicability of management methods. Shepherd notes that managerialism is based on a perception of rationality based on scientific method, and references Klikauer who defines the presentation of managerialism as value neutral and based on unquestioned, common-sense truths (Shepherd 2018:1673). Klkauer also notes that managerialism is about a "managerial-engineering approach to societal problems that have been converted into technicalities" (Klikauer 2015:1107).

Grint and Woolgar contend that through this lens technology is "…construed as the root determinant for either good (technophilia, utopia and hype) or evil (technophobia and dystopia)" (Grint and Woolgar 1997:67). The implication is that technology in its material form is held responsible for both good and bad outcomes, irrespective of its social relations. This limits study to the emergent properties of the machine rather than its systemic relations and consequences. It also provides means to obfuscate the ways in which human interests effect technology and its applications. Like Grint and Woolgar, Walton notes that determinism acts as a normative agent which attributes technology so much power that it can dictate human behaviours (Walton 2019:9–10). Bimber elaborates on normative determinism and suggests it indicates the extent to which human society has relinquished control over technology, replacing ethical norms with technical goals of efficiency and productivity (Bimber 1990:337). He also suggests a theory of nomological determinism, which denies any role for humans in the course of future history, driven instead by artifacts of technology. Thus, determinist preferences in defence cultures provide sets of guiderails around technology that are often unacknowledged but limit how it can be interrogated. This perspective on technology and the epistemological framework which shapes which information and data participates in analysis is the result of a series of historic choices and preferences, to which there are alternatives. The limitations to this framework should be obvious: It does not allow for the agency of users, the complexity of the world and the socio-cultural context of technology to be expressed. If determinism replaces values with those of efficiency and productivity, the sensible question cannot be asked whether efficiency equals efficacy, or whether efficiency is harmful if at the expense of other characteristics and consequences of machines.

Frameworks for trust in technology often betray institutional determinist preferences. As Reinhardt notes in her meta-analysis of ethics and trust literature concerned with AI, "most guidelines are based on an instrumental understanding of trust: trust is described as something that is a precondition to achieve other things" (Reinhardt 2023:737). *Instrumentalised* trust can be equated to pursuit of organisational *determinist* preferences because both are used to justify the normative technological values described by Grint and Woolgar, Walton and Bimber. Where ethical values have been replaced with the pursuit of efficiency and productivity, trust is instrumentalised in its pursuit. The analysis of Yang & Wibowo explicitly suggests that trust building frameworks are designed to produce cognitive and affective change in users, or changes to perceptions, opinions, beliefs, and emotions (Yang and Wibowo 2022:2068), which highlights the idea that trust is an organisational goal rather than a concept containing inherent value. It should be stressed that pursuit of efficiency and organisational goals are not problematic by themselves. However, when trust is instrumentalised as part of a process of optimisation the agency of the user is being neglected and leaves out the possibility that their individual experience with any system can improve its overall capability. Instrumentalisation renders the user as invisible to the generation of efficiency. As quoted by Brown in relation to trust in healthcare services "instrumental trust places an emphasis on the visible performance of the system and tangible experience at the access points to the extent that emotion work and affective interaction are completely overlooked" (Brown 2008:358). Brown also quotes Habermas to highlight that conceptualising the healthcare system as highly rational, numerical and science-based leaves the concern of the patient as "'the pollutants, the sewage of emotionality [which] are filtered off' and where economic/risk-based rationality 'becomes the sole admissible value'" (Brown 2008:359). Instrumental attitudes towards trust are reductive and frame a contest between the trustor who needs to be brought into line with the trustee. As Reinhardt states: "[a] lack of trust is dominantly seen as something to be overcome" (Reinhardt 2023:737). Arguably, therefore, the trustor in this *determinist*, *instrumentalised* view of trust in technology does not have real agency to trust or not to trust.

In the context of a determinist frame for technology, instrumental trust is understood through a prism of the material and displayed behaviours of the user and the technology under study. Some examples of this epistemological framing of technology and trust relations can be found in meta-analyses of the subject. Schaefer's analysis of human-robot trust relations refers to a conceptual framework of factors derived from previous taxonomies built on analysis of interpersonal trust including Muir and Lee & See. Her taxonomy draws from these and breaks down trust factors into three specific categories: *human factors*, *partner or technology factors* and *environment factors* (Schaefer et al. 2016:378). Kaplan also breaks down her taxonomy of human-AI trust into *human (trustor)*, *AI (trustee)* and *contextual factors* (Kaplan et al. 2023:339). Kaplan has worked extensively with Hancock and derives much of her taxonomy from their previous works on robots (see Hancock et al. 2021). Recalling again the capacity of determinist analysis to provide justificatory and normative frameworks for technology, and instrumentalised trust as a means to an optimised end:

> *…we can observe a certain one-sidedness in the guidelines regarding the idea of how trust is established. The focus is clearly on the side of those who have an interest in building trust. Trust is very strongly portrayed as something that one can bring about, that needs to be improved, maintained, earned, and gained: the dominant envisioned actor of the trust game is the trustee. When reading the guidelines, it sometimes appears as if bringing about trust were entirely under control of the trustee. The role of the trustor is not sufficiently reflected* (Reinhardt 2023:738).

Reinhardt is observing that instrumental trust is leveraged to justify and normalise the use of AI applications. This is more than just a philosophical point as an instrumental trust skirts ethical questions around the agency of its users not to trust or structural questions which might benefit from a more sceptical or distrustful attitude towards AI. Reframing the lens through which technology is understood is therefore important as a component of building more effective trust around AI. Some of the meta-analyses of trust reflect more complex understanding. Yang & Wibowo break their taxonomy of human-AI trust into categories of *technology*, *organisational*, *context-related*, *social*, and *user-related* factors. Reinhardt breaks down her paper into *factors affecting knowledge of trust* itself, *trustors*, *trustees*, and a wider category of *factors making AI trustworthy*. Each of these studies relates to the antecedents of trust, meaning the factors that aggregated together result in greater levels of trust in the technology under study. Their meta-analytic methodology indicates there is at least a degree of agreement that some factors are recurrent where academic literature attempts to address trust in technology. Schaefer and Kaplan overlap to a significant degree, and draw on similar literature:

- User, or human factors across these meta-analyses suggest competency, understanding, expertise, experience, workload, demographics, comfort, and attitudes towards AI are all significant. Likewise, personality traits and propensity to trust also matter along with stress and fatigue.
- Technology factors include dependability, performance, behaviour, predictability, reliability, as well as personality, anthropomorphism, appearance, communication, level of automation, reputation, and transparency.
- Contextual or environmental factors include the composition of a team, cultural and social impact, in-group membership, mental models, risk, task, complexity, and context as well as the physical environment.

These factors are intended to be metrified, using some relative understanding to define whether a user-machine relationship is better or worse than one which is believed to generate effective trust. In a framework of understanding such as Operational Research this allows organisations to pursue improvements or optimise the relationship between the user and technology for better trust relations. Even the more complex taxonomy produced by Yang & Wibowo is drawn from an analysis of different behavioural theories, and theoretical frameworks as well as trust antecedents. The taxonomy drawn from this analysis breaks out the contextual factors of Schaefer and Kaplan into organisational, social, and context-related factors:

- Organisational factors are related to compliance with social norms and regulations and the reputation of the organisation.
- Social factors in this context relate to the compliance of the technology with social norms and cultural standards, including integrity and manners, overlapping with perceptions.
- Contextual factors are that of the specific application of technology, including perceptions by users of its utilitarian value and their enjoyment of it. Given factors can vary by application, this can affect the role played by other factors related to trust in technology.

These factors are an important contribution to understanding the structural factors which can affect trust but are presented as factors to be measured and optimised irrespective of the input of the user, and irrespective of the presentation of the technology as a unitary actor. Instrumentalised trust indicates that the organisation believes it can alter the conditions under which technology is understood and thus engender through proximate measures. The different meta-analyses seem to indicate that complex knowledge of the system under study, and the user's relationship with the system are only tangentially relevant. The implication is that there is a power imbalance which affects expectations of the user. Grint and Woolgar wrote of a concern:

> *…with the particular regime of truth which surrounds, upholds, impales and represents technology. Histories which represent themselves as the truth are often the histories of the victor. Thus, we are faced with representations of technology, not reflections of technology. A reflection implies the truth, a representation implies a truth. Similarly, our knowledge of technology–which also represents itself as the truth–is knowledge constructed by the powerful, not by the weak; and, equally significant, by the collective, not the individual* (Grint and Woolgar 1997:32).

They are contending that the way technology is understood should not be reduced to an essentialist account of its properties, because of the way human beings construct a social world around themselves and between each other. Reinhardt is acknowledging this by observing that instrumental trust is brought about through power relations between the trustor and the trustee. Grint and Woolgar also wrote about the way in which the users of computers are understood and concluded that the capacity of a computer can be "construed as a struggle to configure (that is, to define, enable and constrain) the user" by different parts of the design and production community (Grint and Woolgar 1997:73). Thus, it can be argued that the user, or the trustor in their relationship with technology is subject to more than just the power of the machine itself but also the relations around it. A relational understanding of trust informed by assemblages, or similar analytical frame considers the role of the technology, the trustee, and the trustor and importantly allows them agency to trust or not to trust. This approach might seek to interrogate more closely why trust is required and who or what requires trust, the consequences of misplaced trust or over trust in a tool or application.

3.2 Anti-Positivist Epistemologies and Trust

There are a significant number of authors and institutions that contest a determinist framing of technology and suggest alternatives. The academic field of Science and Technology Studies (STS) is a home for such scholarly discourse closely related to works critical of the natural science method. STS is often used as shorthand for a research field which incorporates the work of those specifically seeking to critique determinist understanding of technology as well as others for whom technology is only part of their interest in social systems. This work often starts with a critique of the assumption that artifacts have essential characteristics understandable through measurement and observation and are often strongly anti-positivist. A popular framework for analyses of technology from within this field of study is Deleuzian assemblage theory (see De Landa 2016; Deleuze, Guattari, and Massumi 2013; Nail 2017). It allows the capability and capacity of technology to be understood as a dynamic and constantly evolving relational negotiation between entities. These entities can be humans, other technologies, social and organisational influences, biases, or assumptions. Lisle describes assemblage theory as an analysis which foregrounds the relationship between entities that entangle together at any moment rather than trying to derive essential characteristics of the technology (Lisle 2021:439). To put this more plainly- technology comprises a field of interacting objects, from people and components to ideology, assumption, and bias. From this constant interaction the capability of the technology itself emerges, but sometimes can by understood differently to different communities.

This is where *trust* can be differentiated from *reliance*, and where definitional specificity of technology is important. Trust has already been defined in this paper as the potential for betrayal when a peer fails in their expected duty and is thus more consequential, whilst *reliance* results in disappointment when a tool fails in its purpose. A question to be asked might therefore be whether current AI technology warrants trust as a peer should, or whether the trustee is the web of sociotechnical actors that resulted in its delivery. If this is the case then the factors described in the meta-analysis above are measures of reliance, whilst the question of trust is left open because reliance is more of a technical measure than ephemeral and very human-centred trust. Thus, questions concerning trust in AI must begin by interrogating what constitutes the technology to be trusted, and whether it constitutes a peer. Related to this could be asked at what threshold of technological capability must reliance become trust. However, Sutrop notes:

> *…it may well be that when we speak about trust in AI, in reality we are speaking about trust or distrust of individuals and institutions who are responsible for developing, deploying and using AI. In order to avoid confusion, we should make the object of our attitude of trust clearer* (Sutrop 2019:512).

Arguably, even where humans rely on AI rather than being required to trust it, they must still trust the individuals and institutions that are responsible for its development. Further questions of trust in AI must start with those above by specifying the technology itself, then differentiating what should be trusted and what must be relied on, and then locating where trust must be built. This echoes the point made by Suchman that “AI can be defined as a sign invested with social, political and economic capital and with performative effects that serve the interests of those with stakes in the field” (Suchman 2023:3). Treating the capability of technology as being constructed from an assemblage of interests, individuals and institutions enables more than just a checklist of factors. It enables differentiation between what should be a trusted party and what must merely be relied upon.

In addition to locating and better defining the subject of trust and the technology itself, there is a need to take the agency of the trustor more seriously to prevent their instrumentalisation. Both Sutrop and Reinhardt observe that trust is an ambivalent concept that is to do with uncertainty and with vulnerability (Reinhardt 2023:739; Sutrop 2019:503). Ambivalent trust recognises the inherent risk of uncertainty and vulnerability, and that outcomes can be both positive and negative (Reinhardt 2023:739). Recognising both ambivalence and the agency of the trustor are therefore interleaved concepts, because to recognise the agency of the trustor is to accept that they may view technology and its capacities ambivalently. Distrust and scepticism points at a landscape of more indefinable reasons for distrust that must be explored. The reference to Sutrop above highlights that for all the instrumental, technical reasons why trust can be built, there may be a host of structural reasons why trust and distrust are built.

The assemblage was suggested earlier as a contrasting theoretical framework to a determinist, instrumental approach to understanding trust in AI. It is a relational way of understanding technology. It offers the possibility of overcoming the limitations of a treating AI as a unitary actor by interrogating more closely the technology itself and the limits to instrumentalised trust by interrogating more closely different facets of human-technical interaction. By overcoming these limitations, it also offers a framework within which ethics can be better situated alongside more technical factors affecting trust. An assemblage based, relational or socio-technical approach to understanding technology relies on reconstructing technology in its complexity and in its context. From the perspective of AI, this would mean demystifying it by describing it into a construction of different actors, processes, components, and algorithms from which actions emerge. An example of this might be differentiating the decision-making algorithms in a self-driving car from its suite of sensors, and again from the motive components that make it a vehicle. From this can be build the series of interrelations that comprise the development, maintenance, and use of the technology. By looking at technology and trust as a relational problem, the questions above about specificity and locating can be addressed, and the agency of the trustor can also be situated. Building a socio-technical network of relations also enables trust in institutions to be fully explored. If the technology itself is not considered a peer, then the network of relations enables trust to be located appropriately amongst developers and other structural causes of trust and mistrust.

3.3 Theoretical Framework for Case Study Research

Müller and Schurr describe the assemblage as “a collection of relations between heterogeneous entities to work together for some time” (Müller and Schurr 2016:219). Methodologically, the assemblage captures a moment in which sociotechnical relations coalesce, defined through the capacity of those relations to act on others in a particular way. The assemblage is a flat construction that does not privilege hierarchy or physical distance. Rather it captures interactions between components occurring in a given moment, which might manifest differently from the acknowledged social and technical structure of an institution. One of the powerful descriptive dimensions of the assemblage is its ability to define what stabilises a given socio-technical moment, and through this identify how the power of social relations is enacted. Thus, an assemblage could be used to describe the multiplicity of actors and relations which define the moment of making a phone call, or to describe the capability of an unmanned vehicle as it is demonstrated during the moment of experimentation. Not only does it show the technical relations, but also the power relations which constrain the ability of users to act on the technology and the technology to act on users. For the purpose of trust building research, it will allow how technology functions in the moment that users actively trust or distrust its capacities to be described. To operationalise this theoretical method is to draw extensively on descriptive information. However, there is no objective means to describe any sociotechnical structure. Rather, it can only be a creation based on the perspective of the observer. Thus, any assemblage to be used as a descriptive tool will be subject to reinterpretation by different stakeholders related to the technology. Again, this is helpful when examining fundamentally subjective questions such as trust, as it builds from the perspective of the individual whom the research is trying to draw understanding from.

3.3.1 Operationalisation of Assemblage in Trust Building Research

This research seeks to gather empirical evidence which will allow generalised insights to be drawn in search of guidelines for trust building between users and AI applications. It will do so by drawing on an operationalised version of assemblage theory (see Nail 2017). To create a series of generalised insights around trust building, case studies will be built up using the following process:

1) Background Research: Gathering requirements documents, descriptive documents, details of use cases, experimental designs, results of experiments and other long form information around the AI application under study. Identify event or moment from which to draw assemblage.

2) Identify Actants: From background research components is drawn the actants from which the AI application is built, in the context of the event or moment in which it is being described. Actants include the material technology in its relationships with other pieces of technology as well as people and related socio-cultural beliefs which affect how the technology was built and deployed.

3) Sociotechnical Construction: Articulate as thick description and diagram how the different identified components of the technology link together as an assemblage.

4) Conduct interviews: Interview users and other key stakeholders to understand how trust is conceptualised by them in relation to the AI application under study. Each interview will be based discursive review of the initial assemblage using prompting questions. By working through the initial assemblage, the user can subjectively observe how they saw the technology differently from their perspective, and within that framework point to trusted / reliable or distrusted / unreliable actants and relationships.

5) Write up: From interviews articulate the nature of trust against different actants and how this varies between different users and stakeholders (if applicable).

Additionally, as referred to in the introduction to this paper Suchman identifies AI as a 'floating signifier' which belies a larger field of connected technologies. She notes that the analysis of AI often fails to specify its components, confusing treatment of a field of activity with that of a unitary actor posing existential risks (Suchman 2023:4). Suchman is pointing at a determinist understanding of a theoretical unitary AI actor. Recalling Grint and Woolgar from above, AI is implicitly understood as a potential root determinant for good or evil, and yet AI is not a unitary actor. Suchman is suggesting that by demystifying AI though exposing its assembled components the source of this theoretical power can be better understood.

4.0 CONCLUSION

The empirical research suggested here will create assemblage-based case studies as a complement to insights drawn from a living literature survey. They will allow trust to be situated in the specific context of the UK defence establishment and provide new knowledge for the field of trust building research. This approach to understanding trust relations has the potential to overcome limitations imposed by a factorial approach to trust which is defined by the natural scientific method. Although the quantification of performance metrics allows a measure of understanding of how users can expect to rely on the AI based tool they are deploying, this STS informed approach allows trust to be situated amongst the organisation and institution in which the user works. This knowledge allows further organisational change to address issues of mistrust, distrust or overtrust. The next steps for this programme of research would be to develop a research instrument and ensure that the relevant ethical clearance processes are undertaken before reaching out to possible AI application users for interview.

5.0 BIBLIOGRAPHY

[1] Baier, Annette. 1986. 'Trust and Antitrust'. Ethics 96(2):231–60. doi: 10.1086/292745.

[2] Bimber, Bruce. 1990. 'Karl Marx and the Three Faces of Technological Determinism'. Social Studies of Science 20(2):333–51. doi: 10.1177/030631290020002006.

[3] Brown, Patrick R. 2008. 'Trusting in the New NHS: Instrumental versus Communicative Action'. Sociology of Health & Illness 30(3):349–63. doi: 10.1111/j.1467-9566.2007.01065.x.

[4] De Landa, Manuel. 2016. Assemblage Theory. Edinburgh: Edinburgh University Press.

[5] Deleuze, Gilles, Félix Guattari, and Brian Massumi. 2013. A Thousand Plateaus : Capitalism and Schizophrenia. London: Bloomsbury.

[6] Grint, Keith, and Steve Woolgar. 1992. 'Computers, Guns, and Roses: What's Social about Being Shot?' Science, Technology, & Human Values 17(3):366–80. doi: 10.1177/016224399201700306.

[7] Grint, Keith, and Steve Woolgar. 1997. The Machine at Work : Technology, Work, and Organization. Cambridge, UK: Polity Press.

[8] Hagendorff, Thilo. 2020. 'The Ethics of AI Ethics: An Evaluation of Guidelines'. Minds and Machines 30(1):99–120. doi: 10.1007/s11023-020-09517-8.

[9] Hancock, P. A., Theresa T. Kessler, Alexandra D. Kaplan, John C. Brill, and James L. Szalma. 2021. 'Evolving Trust in Robots: Specification Through Sequential and Comparative Meta-Analyses'. Human Factors: The Journal of the Human Factors and Ergonomics Society 63(7):1196–1229. doi: 10.1177/0018720820922080.

[10] Kaplan, Alexandra D., Theresa T. Kessler, J. Christopher Brill, and P. A. Hancock. 2023. 'Trust in Artificial Intelligence: Meta-Analytic Findings'. Human Factors: The Journal of the Human Factors and Ergonomics Society 65(2):337–59. doi: 10.1177/00187208211013988.

[11] Klikauer, Thomas. 2015. 'What Is Managerialism?' Critical Sociology 4:1103–19.

[12] Lisle, Debbie. 2021. 'A Speculative Lexicon of Entanglement'. Millennium: Journal of International Studies 49(3):435–61. doi: 10.1177/03058298211021919.

[13] McKnight, D. Harrison, and Norman L. Chervany. n.d. 'What Is Trust? A Conceptual Analysis and An Interdisciplinary Model'.

[14] Müller, Martin, and Carolin Schurr. 2016. 'Assemblage Thinking and Actor-network Theory: Conjunctions, Disjunctions, Cross-fertilisations'. Transactions of the Institute of British Geographers 41(3):217–29. doi: 10.1111/tran.12117.

[15] Nail, Thomas. 2017. 'What Is an Assemblage?' SubStance 46(1):21–37. doi: 10.3368/ss.46.1.21.

[16] Reinhardt, Karoline. 2023. 'Trust and Trustworthiness in AI Ethics'. AI and Ethics 3(3):735–44. doi: 10.1007/s43681-022-00200-5.

[17] Schaefer, Kristin E., Jessie Y. C. Chen, James L. Szalma, and P. A. Hancock. 2016. 'A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems'. Human Factors: The Journal of the Human Factors and Ergonomics Society 58(3):377–400. doi: 10.1177/0018720816634228.

[18] Shepherd, Sue. 2018. 'Managerialism: An Ideal Type'. Studies in Higher Education 43(9):1668–78. doi: 10.1080/03075079.2017.1281239.

[19] Suchman, Lucy. 2023. 'The Uncontroversial "Thingness" of AI'. Big Data & Society 10(2):20539517231206794. doi: 10.1177/20539517231206794.

[20] Sutrop, M. 2019. 'SHOULD WE TRUST ARTIFICIAL INTELLIGENCE?' Trames. Journal of the Humanities and Social Sciences 23(4):499. doi: 10.3176/tr.2019.4.07.

[21] Walton, Steven A. 2019. 'Technological Determinism(s) and the Study of War'. Vulcan 7(1):4–18. doi: 10.1163/22134603-00701003.

[22] Yang, Rongbin, and Santoso Wibowo. 2022. 'User Trust in Artificial Intelligence: A Comprehensive Conceptual Framework'. Electronic Markets 32(4):2053–77. doi: 10.1007/s12525-022-00592-6.