HFM-300 Symposium on Human Autonomy Teaming (HAT)
Trust Me I'm Artificial Intelligence

Fiona Butcher
Defence Science and Technology Laboratory (Dstl)
UNITED KINGDOM

fdbutcher@dstl.gov.uk

The contents of this document should not be interpreted as representing the views of the UK MOD, nor should it be assumed that they reflect any current or future UK MOD policy. The information contained in this document cannot supersede any statutory or contractual requirements or liabilities and is offered without prejudice or commitment.

ABSTRACT

The presentation will explain an ongoing research study that addresses the symposium's objective to consider the human factors of trust in autonomous systems and, specifically, the topic of the ethical and social aspects of HAT. The research aims to validate a conceptual model that has been developed to understand how a person's intention to initially trust an artificial intelligence (AI) system may be influenced by a) prior beliefs and attitudes about AI technology b) social influences and organisational norms and c) a belief in the level of control an individual has when deciding to initially trust an AI technology. Using an experimental design, an on-line survey will be administered to consenting UK military and Defence civilian participants who will complete measures that assess their decision making preferences, knowledge and attitude towards technology in general, as well as AI specifically. In addition, the participants will provide responses to an imaginary scenario to assess why and when they would trust an AI system. The scenario topic reflects a typical example of the trust challenge that humans may face when using new AI decision aides to support nuanced human decision making. The presentation will provide an explanation of the conceptual model, the empirical research method and preliminary research findings.

1.0 BACKGROUND

This empirical study is part of Dstl's Autonomous Systems research programme and sits within a project that is exploring the social, legal and ethical considerations of MoD using these systems in the future. The research aims to validate a conceptual model that has been developed to understand how a person's intention to initially trust an artificial intelligence (AI) system may be influenced by a) prior beliefs and attitudes about AI technology b) social influences and organisational norms and c) a belief in the level of control an individual has when deciding to initially trust an AI technology.

As part of an on-line survey, consenting UK participants will complete measures that assess their decision making preferences, knowledge and attitude towards technology in general, as well as AI specifically. In addition, the participants will be provided with information about how a proposed new AI system will be used in Defence and a scenario in which they will be asked to use the system before then being asked a series of questions about why and when they will trust the AI system.

The scenario is based on a recruitment AI system as this is a typical example of the trust challenge that humans may face when using new AI decision aides to support nuanced human decision making. This topic has been chosen as recruitment is considered a strategic priority for MOD and there are a number of commercial providers who are beginning to offer AI systems to aid recruitment decisions (Jeffery, 2017).

It is hoped that the model being tested in this study will provide evidence to inform future MOD assurance policy as well as military training and education policy.

1.1 What is Artificial Intelligence (AI)?

AI has been listed as one of the 'Eight Great Technologies' (GoScience, 2017, p. 5) for the future economy and is defined as "more than the simple automation of existing processes: it involves, to greater or lesser degrees, setting an outcome and letting a computer program find its own way there." (GoScience, 2016, p.5). Rather than an automated system following prescribed routines the AI is coded with probabilistic equations known as algorithms that autonomously seek out different possible routes to reach a goal.

1.2 Why does AI matter to Defence?

The Defence Concepts and Doctrine Centre describes a future in which MOD personnel will be required to make quick decisions in an increasingly complex and cluttered operational environment, managing large volumes of ever changing information (UK MOD, 2014a). In this new environment it is anticipated that Defence will continue to be a significant recruiter of people in the UK. To meet the demands of this new operating environment it is predicted that over the next two decades military operations will increasingly rely on AI and advanced automated systems as part of military planning and operational decision-making processes (UK MOD, 2014b). Figure 1 below illustrates how it is anticipated future military commanders' decision making will be augmented by the use of AI technologies, automating elements of the decision making process. One of the challenges with using this technology will be to ensure that the way the information has been filtered and processed by the autonomous system is sufficiently trusted by the decision-maker.

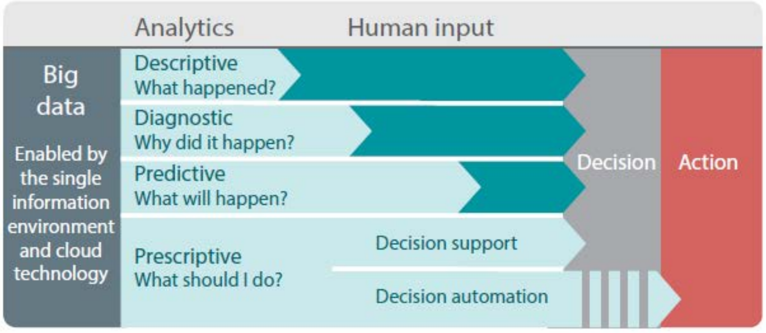

Figure 1: Types of analytics capability[1]

1.3 How is trust in AI being defined?

There is no single agreed definition of trust, however when the technology is an AI that is learning and adapting it is suggested that the interaction becomes more like a human-human interaction. Therefore, interpersonal trust theories (Lewicki & Bunker, 1995; Lewicki, McAllister, & Bies, 1998) have been drawn upon in this study to define 'trust'. These theories suggest a person may initially trust in a new AI system due to affective and social factors. When to trust will be influenced by the context in which the AI system is being used (Hoffman, 2017). The proposition that these factors influence initial trust is supported by Li, Hess, & Valacich, (2008) in which they found when a new information system is proposed initial trust in the

[1] See: UK MOD (2017), Development Concepts Doctrine Centre, Joint Concept Note 2/17, Future of Command and Control www.gov.uk/mod/dcdc

technology is influenced by the reputation of the technology and the belief norms about technology in an organisation.

2.0 STUDY

Following an exploratory literature review an Initial Trust in AI Model (ITAIM) has been developed for this study and is shown in Figure 2. The model is an adaptation of Ajzen's Theory of Planned Behaviour (TPB) (Ajzen, 1991; Ajzen, 2011; Cooke & French, 2011; Yan et al., 2013) linking a person's beliefs with an intention to perform a particular behaviour. The model focuses on initial trust as AI is a new technology that is just beginning to be introduced into military organisations. It will be for future research to investigate how trust develops and changes over time once this new technology has become more established.

The ITAIM describes an intention to trust a new AI system based upon three factors: 1) prior beliefs and attitudes about AI technology, 2) the influence of organisational social norms (this will include the reputation of the AI system), and 3) the perceived level of control in deciding whether to trust (this will include decision making style preferences).

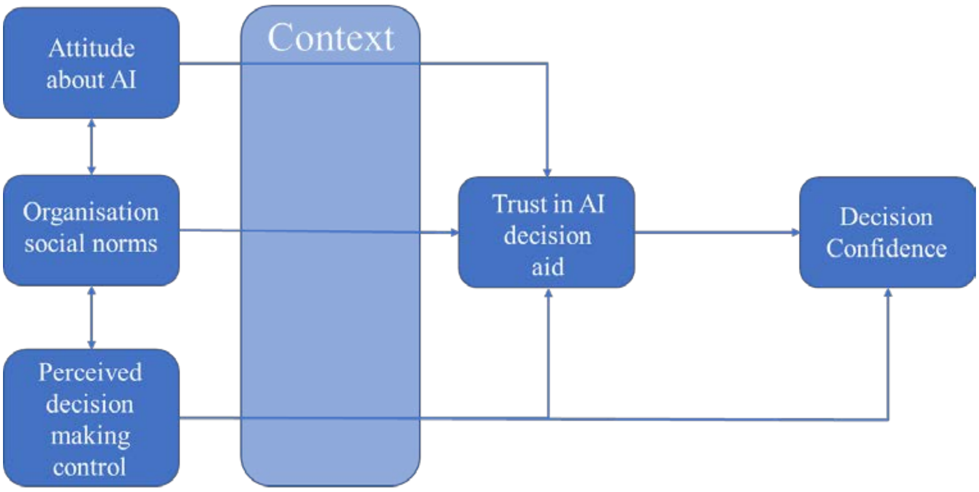

Figure 2: Initial Trust in AI Model (ITAIM)

In applying the model this study aims to answer the following research questions:

- Can knowledge, beliefs and attitudes about AI influence a person's level of initial trust in an AI decision aid?
- Can decision-making preferences (such as a preference for maximum information before making a decision) influence the level of initial trust a person has in an AI system?
- Does the level of initial trust in an AI decision aid influence the level of confidence a person has in their decision to trust?
- Does the reputation of the AI tool influence a person's initial trust in AI tools?

2.1 Overall Methodology

This study will carry out a between-subjects randomised control experiment. The experiment will be completed on-line using the SurveyMonkey platform. Consenting MoD participants will be asked to complete pre-test measures that capture their (a) biographical data (b) decision making preferences and (c) their knowledge and attitude towards AI. Post-test measures will assess (a) attitudes and beliefs about the AI system described in the scenario (b) level of confidence in their decision to trust the AI system and (c) levels of trust in technology generally. These self-report measures will be combined for administration purposes into single pre and post-test questionnaires.

The dependant variables are the level of initial trust in AI system and confidence in the decision to initially trust. The independent variable is information on the reputation of the tool. The covariates are: knowledge and attitudes about AI, and decision-making preferences.

Listed below are potential hypotheses that will be tested:

H0(null): There is no effect on an individual's intention to initial trust advice from an AI decision system as a result of the individual's beliefs, attitudes or decision making preferences, nor the reputation of the AI system.

H1: Those individuals who have positive attitudes and beliefs about AI technology are likely to initially trust AI decision aids.

H2: Those who seek less information will perceive themselves as having higher levels of control in deciding to initially trust.

H3: Individuals who are likely to initially trust an AI tool are also more confident in the decisions they make using an AI decision aid.

H4: Individuals with high initial trust will also be more confident in their decision to trust.

H5: The reputation of the AI system will influence an individual's initial trust in the system.

An overview of the experimental process is shown in Figure 3 below. The design has been piloted with Dstl employees to test the online survey items suitability before being administered.

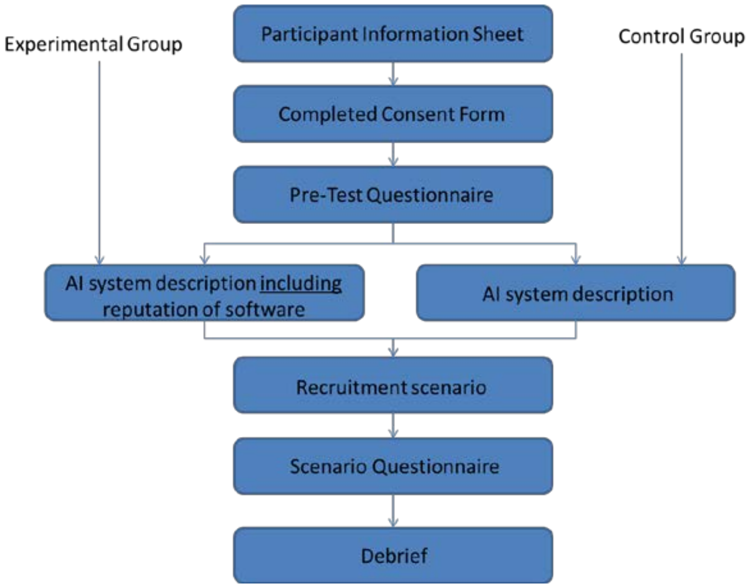

Figure 3: Overview of experimental design for data collection

2.2 Study Procedure

Participants will be recruited from those who have either recently completed or are part of the current intake of 300 personnel on the Advanced Command Staff Course (ACSC) course delivered by the Defence Academy at Shrivenham. This group of participants have a high degree of homogeneity and is a post-graduate level course for mid-seniority military and civilian personnel who will have similar levels of experience of working in Defence. By choosing this group it is hoped to control for factors such as experience and rank.

Participants will be randomly allocated to either the experimental or control group and provided with a unique identifier so that the numbers in each condition can be monitored to ensure that they are equal. All participant information and questionnaire responses will be collected via SurveyMonkey. All consenting participants will be asked to complete the following online questionnaires:

Demographic questionnaire. Participants will be requested to complete a series of demographic questions that will be used to ascertain the extent to which the sample is representative of a mid-seniority military and civilian Defence population. The demographic questions will ask participants to report their gender, whether they are a UK citizen, educational level, role (military or civilian), rank and length of service.

Decision-making tendencies questionnaire. Participants will also be asked to complete a questionnaire on their decision-making tendencies to assess whether their tendencies influence their belief in the level of self-control to decide to trust AI.

Schwartz (2000) suggests that some individuals consistently attempt to seek out the best solution before making a decision (which demands an exhaustive search of the options), while others consistently attempt to find a solution that is satisfactory or good enough given their standards (which can be met by a non-exhaustive search). To measure the degree to which a decision-maker is a 'maximizer' versus a 'satisficer' Schwartz et al. (2002) developed a 13-item Maximization Scale, with maximizing and satisficing at polar opposites of a continuum.

More recently, Misuraca et al. (2015) developed a new measure that incorporates Schwartz et al.'s items and this new measure of decision-making tendencies has been shown to have good construct validity across six dimensions. The 29-item Decision Making Tendency Inventory (DMTI) measures the tendency to maximize, to satisfice, and a third new construct: to 'minimize', which is the tendency to minimize the amount of resources in order to get the minimum of the possible results. Participants respond to each item using a behaviourally anchored 7-point, Likert-type scale (1 = completely disagree,7 = completely agree).

The DMTI questionnaire items will be used and will extend the DMTI research further by investigating whether there is a correlation between the tendency to maximize, satisfice and minimize and actual decision making behaviour (i.e. intention to trust an AI system).

Survey about AI and Machine Learning perceptions. This survey has been adapted from the Ipsos MORI 2017 survey used in the UK public research and engagement conducted on behalf of the Royal Society . The survey was developed in accordance with the requirements of the international quality standard for Market Research, ISO 20252:2012 and the task for the research was to create an evidence base about public perceptions of the potential opportunities and risks of machine learning. By administering items from this survey it will be possible to begin to explore the perceptions of the ACSC participants compared to the UK public.

Once the questionnaires are completed participants will read a short scenario. A recruitment scenario has been chosen as this is a typical example of the trust challenge that humans may face when using new AI decision aides to support nuanced human decision making.

The experimental group scenario description includes persuading statements about the reputation of the AI recruitment system and includes the following words:

…with more than 20 years' experience … Tai systems is considered an ethical system and is used by some of the world's leading companies and in 2017 was listed as one of the top five recruitment software developers in the UK…. reducing the time spent by companies in reviewing candidates by up to 70%...Recruitment policymakers believe we can save time and money by using such technology to accurately find candidates who meet our exact job requirements at an early stage.

The statements outlined above have been designed around three of Cialdini's principles of persuasion; which are scarcity, authority and consensus. After observing real-world influence techniques and reviewing related empirical research, Cialdini (2009) has defined six principles of social influence routinely employed to persuade potential consumers. Cialdini's research shows that people tend to want more of the things they can have less of (e.g. scarcity of time to recruit new personnel), will follow the lead of knowledgeable credible experts (e.g. those software developers who are world leaders in their field) and will also look to the actions of others to determine their own (e.g. the views of other recruitment policymakers).

The **control group scenario description** does not include these persuading statements (see below):

Tai Systems a company that has developed technology that uses Machine Learning (a form of Artificial Intelligence) to help employers to identify the best-suited candidates in the shortest amount of time. A Tai system supplies a broad range of companies that are interested in recruiting new personnel. Tai Systems is on a mission to make recruitment more efficient, engaging and inclusive, applying their technology, helping companies recruit people that will adhere to their values, and thrive in their culture.

Once the participant has read the scenario they will be asked to complete an online questionnaire about the recruitment system. The first seven items in the questionnaire have been developed using Hoffman's 'trust in automation' taxonomy (Hoffman, 2017). This taxonomy is a nuanced classification of human trust in machines as a process that is dynamic over different contexts leading to a variety of trusting relationships. In developing this taxonomy Hoffman has considered the various factors, dimensions and varieties of trust that have been proposed by previous theories and empirical research and recognises the complications and complexities of trust.

The remaining scenario response items are designed to assess the participant's behavioural intentions, attitudes, subjective norms and perceived behavioural control. The design of these items has been informed by Francis, Eccles, Johnston, Walker, Grimshaw, Foy, Kaner, Smith, & Bonetti,(2004) Theory of Planned Behaviour Questionnaires: Manual for Researchers which was developed for researchers throughout the European Union involved in the Research-Based Education and Quality Improvement (ReBEQI) project. The manual has been subjected to comprehensive review and trialling procedures as part of the ReBEQI project and is therefore considered an authoritative guide on the theory of planned behaviour questionnaire development. There is one item (with four behaviourally anchored 7-point, Likert-type scales) that directly measure the participant's attitude towards initially trusting the Tai system. This is followed by four items that measure subjective norms using incomplete sentences that the participant indicates a response to on a 7-point likert-style scale. The final four items are statements with 7-point Likert-style responses that assess the participant's self-efficacy i.e. how confident they are in trusting Tai and how much control they have over whether they can decide whether to trust Tai. Once the participant has completed the questions about the scenario they will be asked about their general attitude and trust in technology.

2.3 Data Analysis Method

A between groups analysis of variance will be conducted to test the individual performance hypotheses with the Independent Variable:

- Information on the reputation of the tool

and Dependent Variables:

- Level of initial trust in the Tai system
- Confidence in the decision to initially trust the Tai system.

An analysis of covariance will be conducted to check for possible confounding effects due to knowledge and attitudes about AI, and decision-making preferences.

A factorial model (ANOVA) will be used to assess the magnitude and significance of fixed effects on a dependent variable. Accordingly, multiple hypotheses will be tested. To avoid the familywise error an adjustment will be made to significance values using Tukey's method. Averages from the questions in the AI trust questionnaire for each participant will be used to generate the dependent variable; between 1 and 7 representing an individual's tendency to trust or reject AI decisions. Prior to this questionnaire the participants will have also answered two questionnaires; firstly, regarding their knowledge and attitude toward AI, and secondly regarding their decision making processes. The first questionnaire, knowledge about AI, will be processed into three distinct outcomes (high, medium, low). The second will be treated similarly to the dependent variable in that an average score between 1 and 7 will be calculated. Following this, participants will be assigned to two groups given differential information, a control and an experimental group. The fixed effects model will include three independent main-effects variables, and all feasible interactions; the experimental group (control or experimental) and the outcomes of the knowledge and attitudes toward AI and decision making questionnaires (categorical and continuous respectively).

Table 1: Data Analysis Plan

Variable Name	Type	Model position
AI trust questionnaire	(1 to 7)	Dependent variable
Experimental group	Categorical (2 levels)	Fixed effect
Knowledge & attitudes	Categorical (3 levels)	Fixed effect
Decision making preference	Categorical (3 levels)	Fixed effect

The exact nature of the multivariate model used will depend upon the distribution of the response variable.

Qualitative data analysis on the free response question 'why?' the participant will or will not trust will be analysed using thematic analysis to identify themes between the two groups of the participants. Braun and Clarke's (2006) approach to thematic analysis will be performed through the process of coding in six phases to create established, meaningful patterns. These phases are: familiarization with data, generating initial codes, searching for themes among codes, reviewing themes, defining and naming themes, and producing the final report

3.0 NEXT STEPS

Data collection is underway and it hoped that preliminary analysis will be completed by early October so that early results can be shared with the HFM-300 HAT Symposium. The trust and social implications for the future development and operationalisation of AI systems that support human decision making and this research will provide MoD with insights that should be considered in the policy, doctrine and use of this technology.

4.0 ACKNOWLEDGEMENTS

The author would like to thank colleagues within Dstl and her supervisor at University of Leicester for numerous spirited discussions that informed this paper. In addition, the paper benefitted significantly from comments received during the Dstl technical review process.

5.0 REFERENCES

[1] Ajzen, I. (1991). The theory of planned behavior. Organizational Behavior and Human Decision Processes, 50, 179–211.

[2] Ajzen, I. (2011). The theory of planned behaviour: Reactions and reflections. Psychology & Health, 26(9), 1113–1127.

[3] Braun, V. & Clarke V. (2006). "Using thematic analysis in psychology". Qualitative Research in Psychology. 3 (2): 93.

[4] Cialdini, R. B. (2009). Influence: Science and practice (5th ed.). Boston: Allyn & Bacon.

[5] Cooke, R., & French, D. P. (2011). The role of context and timeframe in moderating relationships within the theory of planned behaviour. Psychology & Health, 26(9), 1225–1240.

[6] Francis, Eccles, Johnston, Walker, Grimshaw, Foy, Kaner, Smith, & Bonetti,(2004) Constructing questionnaires based on the theory of planned behaviour: A manual for health services researchers. Newcastle upon Tyne, UK: Centre for Health Services Research, University of Newcastle upon Tyne.

[7] GoScience (2016). Artificial intelligence: opportunities and implications for the future of decision making. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/566075/gs-16-19-artificial-intelligence-ai-report.pdf

[8] GoScience (2017). Technology and Innovation Futures 2017. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/584219/technology-innovation-futures-2017.pdf

[9] Hoffman, R.R. (2017). A taxonomy of emergent trusting in the human-machine relationship. In P. Smith & R.R. Hoffman (Eds.) (2017). Cognitive systems engineering: The future for a changing world. Boca Raton, FL: Taylor & Francis.

[10] Jeffery, R., (2018). Don't hire the candidate who presses this key. How algorithms are rewriting the rules of recruitment (And what that means for your job). People Management, December 2017 / January 2018.

[11] Lewicki, R. J., & Bunker, B. B. (1995). Trust in relationships: A model of development and decline. Jossey-Bass.

[12] Lewicki, R. J., McAllister, D. J., & Bies, R. J. (1998). Trust and distrust: New relationships and realities. Academy of Management Review, 23(3), 438–458.

[13] Li, X., Hess, T. J., & Valacich, J. S. (2008). Why do we trust new technology? A study of initial trust formation with organizational information systems. Journal of Strategic Information Systems, 17(1), 39-71.

[14] Misuraca, R, Faraci, P, Gangemi A., Carmeci F.A., Miceli S. (2015). The Decision Making Tendency Inventory: A new measure to assess maximizing, satisficing, and minimizing, Journal of Personality and Individual Differences, 85 (2015), 111–116.

[15] Schwartz, B. (2000). Self-determination: The tyranny of freedom. American Psychologist, 55, 79–88.

[16] Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., & Lehman, D. R. (2002). Maximizing versus satisficing: Happiness is a matter of choice. Journal of Personality and Social Psychology, 83, 1178–1197.

[17] UK MOD (2014a). Defence Concepts and Doctrine Centre. Strategic Trends Programme: Future Operating Environment 2035, (First Issue). Retrived from https://www.gov.uk/government/publications/future-operating-environment-2035

[18] UK MOD (2014b). Defence Concepts and Doctrine Centre. Strategic Trends Programme Global Strategic Trends - Out to 2045 Fifth Edition. Retrieved from https://www.gov.uk/government/publications/global-strategic-trends-out-to-2045

[19] Yan, Z., Dong, Y., Niemi, V., & Yu, G. (2013). Exploring trust of mobile applications based on user behaviors: An empirical study. Journal of Applied Social Psychology, 43(3), 638–659.