

# HOW DO WE UNDERSTAND TRUST?

*Theatre, AI and 'Ludic Technologies'*

*Humanist Perspectives*

*9<sup>th</sup> April, 2021*

# INTRODUCTION

# HUMAN-MACHINE TEAMING

Automation aims to reduce operators' workload, costs and errors while increasing precision. It can help with:

- Improved levels of safety
- Enormous savings
- Operation in safety critical conditions
- More effective use of the equipment.

## However

The role of humans have moved from direct control to supervision – this creates new challenges.

Cooperation is at the core of Human-Machine Teaming – without this, forming an effective team is impossible.

One of the main goals is to establish strong and cooperative relationships in which people find the system credible, acceptable and usable.

A key factor that influences this is **trust**.



# THE ROLE OF TRUST

Trust is one of the key factors that determine whether an operator will choose to use automation:

- Trust is a key component of a socio-technical approach to collaboration
- Trust also impacts the relationship between human and cyber-physical elements, and the formation of hybrid systems
- Trust encourages and ensures effective human-autonomy operations, particularly in dynamic and uncertain environments
- Trust is necessary for the effective implementation of specific roles and responsibilities in a human-autonomous team

A hand is shown placing a white letter block with the letter 'T' on a blue surface. To the left of the hand, four other white letter blocks are already in place, spelling out 'TRUS'. The blocks are arranged in a horizontal line, and the hand is positioned to the right, about to place the final 'T' block to complete the word 'TRUST'.

T R U S T

# TRUST AS A PROCESS

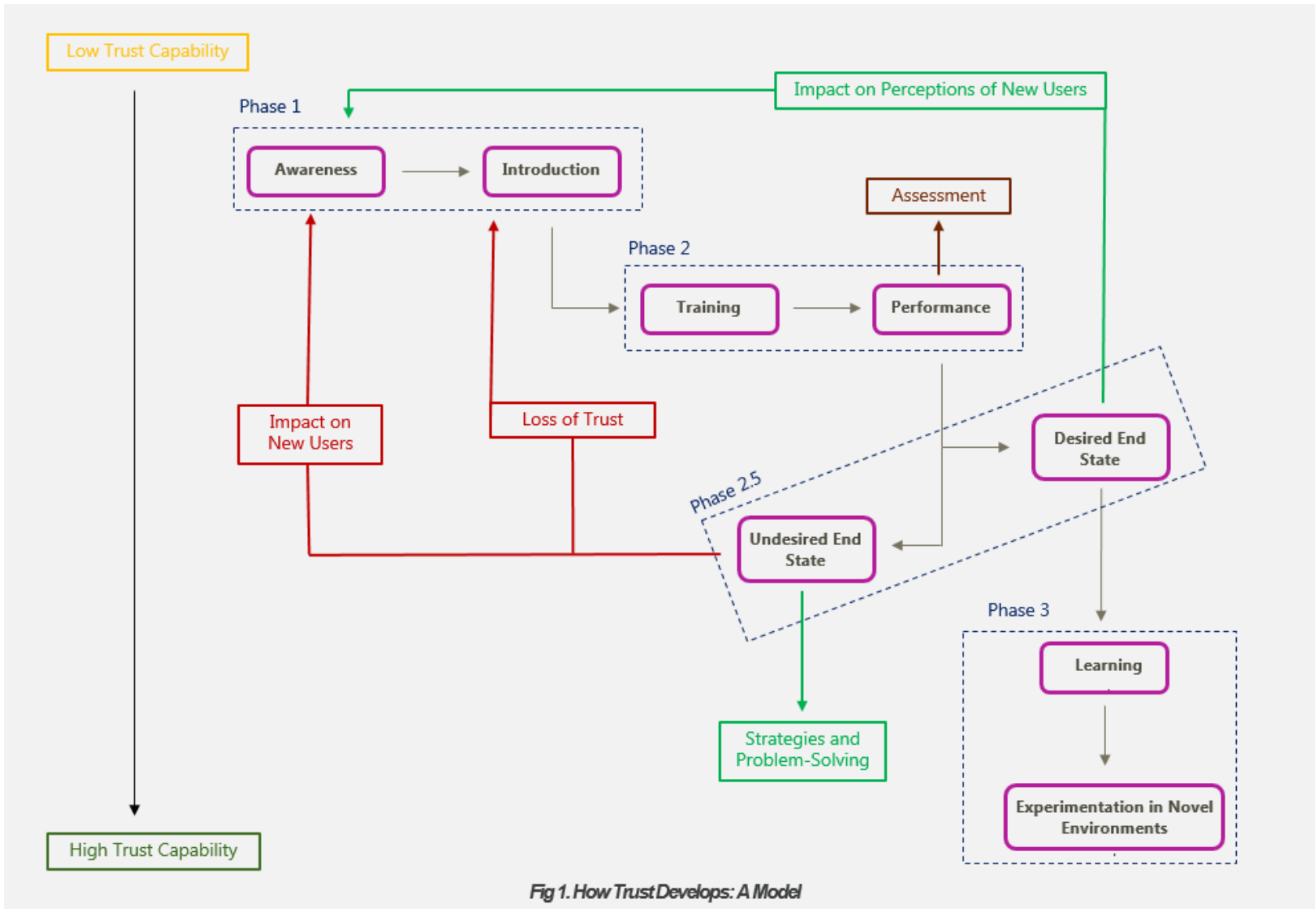


Fig 1. How Trust Develops: A Model

# BASIC DESIGN PRINCIPLES

**Transparency:** the human must be able to understand what the system is doing, and ask questions

**Humans Involved in Decision Making:** the human must be kept in the loop and continue to be involved in the task(s)

**Ease of Interaction:** the AI interface should be easy to engage with.

**Training and Experience:** training regarding the system would help the human work with it better.

**Reputation and Regulation:** the reputation of the manufacturer, familiarity with the technology and an understanding of the context helps with the development of trust.

**In-Group Identity:** trust is more likely to develop if teammates are perceived to be from the same group with common goals and values.

**Rich Communication:** the system will need to be able to engage in rich communication to convey information – the use of natural language increases the chances of its acceptance.

**Agency:** the system should be designed with anthropomorphic features and agency prior to being put in a partnership with a human.

# EXERCISE ONE

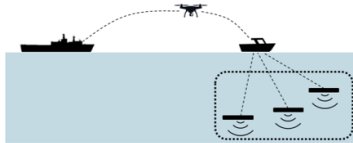
# EXERCISE ONE

Participants were presented with two scenarios:

## SCENARIO 1

### SEABED SURVEY: CONDUCT A TIMELY SURVEY OF THE SEABED IN A DEFINED AREA

Seabed surveys generate a significant amount of data that requires processing - enabling in-stride processing of the data improves time efficiency. Autonomous underwater and surface vehicles can help increase survey coverage, however, size, weight and power constraints limit their processing capabilities. Offloading processing of data from them to a host ship with an autonomous aerial vehicle will help mitigate the risk of losing the vehicle and data collected. Assumptions are that data transfer to/from the vehicle is possible, communications are available and the mission planning systems for the assets are on the host ship.



The questions that this raises include:

How does the (human) mission supervisor maintain accountability for this squad of assets, performing a collaborative task?

How does this change of role (moving the person onshore) change their trust relationship with their technology and their colleagues? Could this drive different behaviours?

How can/should other users of the sea trust this system, as they come across it?

As data is being received from this asset, which we may not be able to see, can we really trust what it is telling us?

## SCENARIO 2

### INTEGRATING INTELLIGENT MACHINES INTO THE HUMAN BODY

Currently, intelligent machines do not have the capacity to reciprocate trust in their human teammates or hold intrinsic values or self-awareness. However, the 2030s will herald a convergence of nanotechnology, biotechnology, information technology, and cognitive science (NBIC) to bring about a wide-spread adoption of technologies designed to help humans to become stronger, smarter, more resilient, and to cultivate new abilities. Humans with enhanced intelligence and enhanced physical abilities may form teams with humans.

Technologies that are predicted to be commonplace by 2030 include: implanted RFID chips for identification and security, exoskeletons, surgical enhancements to our bodies, real-time language translation, augmented vision, smart contact lenses, direct brain interfaces to enhance communication and sensory perception and integration and that may provide the ability to capture a thought and share it instantly.



Their task was to answer the following questions:

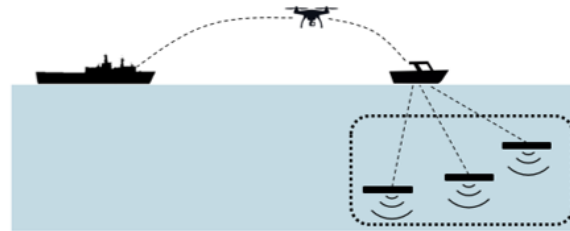
1. What aspects of the situation seem trustworthy?
2. What doesn't seem trustworthy?
3. What are the possible problems – trust or otherwise – that you can foresee with the system?



# SCENARIO 1

## SEABED SURVEY: CONDUCT A TIMELY SURVEY OF THE SEABED IN A DEFINED AREA

Seabed surveys generate a significant amount of data that requires processing - enabling in-stride processing of the data improves time efficiency. Autonomous underwater and surface vehicles can help increase survey coverage, however, size, weight and power constraints limit their processing capabilities. Offloading processing of data from them to a host ship with an autonomous aerial vehicle will help mitigate the risk of losing the vehicle and data collected. Assumptions are that data transfer to/from the vehicle is possible, communications are available and the mission planning systems for the assets are on the host ship.



The questions that this raises include:

How does the (human) mission supervisor maintain accountability for this squad of assets, performing a collaborative task?

How does this change of role (moving the person onshore) change their trust relationship with their technology and their colleagues? Could this drive different behaviours?

How can/should other users of the sea trust this system, as they come across it?

As data is being received from this asset, which we may not be able to see, can we really trust what it is telling us?

## POSSIBLE PROBLEMS

- Collision with other vehicles
- Damage to wildlife/surroundings
- Other “users of the sea” might tamper with equipment
- Impact of weather conditions
- Connectivity issues
- Collection and storage of “unintended” data
- Possible tension between the human and the autonomy – who is the blame for failures assigned to?

## TRUSTWORTHY

- Unity amongst the group according to a hierarchy that relies on communications with off shore host ship.
- Reduced risk by being able to control assets centrally/at surface
- Reduced risk of human error
- Reduced risk of data loss

## NOT TRUSTWORTHY

- Vulnerability to external interests – lose control, hacking, hijacking
- Potential to lose control of assets
- Threat of job loss
- Data corruption – users may not believe the data
- Are assets really where we think they are?

# SCENARIO 2

## INTEGRATING INTELLIGENT MACHINES INTO THE HUMAN BODY

Currently, intelligent machines do not have the capacity to reciprocate trust in their human teammates or hold intrinsic values or self-awareness. However, the 2030s will herald a convergence of nanotechnology, biotechnology, information technology, and cognitive science (NBIC) to bring about a wide-spread adoption of technologies designed to help humans to become stronger, smarter, more resilient, and to cultivate new abilities. Humans with enhanced intelligence and enhanced physical abilities may form teams with humans.

Technologies that are predicted to be commonplace by 2030 include: implanted RFID chips for identification and security, exoskeletons, surgical enhancements to our bodies, real-time language translation, augmented vision, smart contact lenses, direct brain interfaces to enhance communication and sensory perception and integration and that may provide the ability to capture a thought and share it instantly.



## POSSIBLE PROBLEMS

- Technological segregation
- Access to information and personal data – how is it used?
- Who regulates it? How do you know if it's gone too far?
- Ethical questions are not addressed
- Coding biases
- Non-participatory design
- User Agency may be removed as tech can override intent
- Do we really need this?
- Could lead to prejudice and eugenics

## TRUSTWORTHY

- Removing the element of human emotion from certain decisions
- Removal of some bias - biometric data as incontestable evidence
- Technology is quite advanced, extension of existing technologies
- Potential efficiencies/ enhanced decision-making/ higher performance
- People are more aware of trust/security/data issues than they used to be

## NOT TRUSTWORTHY

- Assumption that “enhanced humans” will be “better”
- Accessibility issues
- “The chip made me do it” – more control over humans
- What is the motivation behind providing this tech? Can it be abused?
- Are my thoughts my own? Will someone “hack” my body?
- Could have long-term health impacts

# EXERCISE TWO

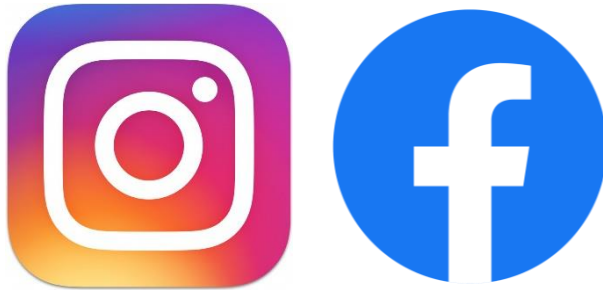
# EXERCISE TWO

Participants had to think about the products and services that they currently use, within the context of trust. They came up with:



# EXERCISE TWO

Two of the options were selected:



**SOCIAL MEDIA**



**YOUR PHONE**

Participants had to use them as possible use cases, in order to answer the following design questions:

- Is there transparent and rich communication from the system to the user?
- Does the human still maintain some control when using the system?
- Is the interface – and the system – easy to use?
- If not, is there a process in place to train users?
- Is there a chance that the system will develop a “bad reputation” – how will that be handled?
- Can the system be considered as being a teammate?
- If not, how can it become a part of the team?

# SOCIAL MEDIA

## TRANSPARENCY

- Are these systems here to serve users or make money?
- Social media sites need to be more transparent with how they collect/use/sell data

## CONTROL

- Risk of social media addiction
- The human is in control of their "settings" but are they in control of their data?
- Responsibility to users
- We "curate" according to the systems' patterns, not our own
- People are blocked when they express "controversial opinions"

## EASY-TO-USE INTERFACE

- Frequent updates with poor user training
- Rapid changes without analysis of user impact
- Social media is not easy for everyone to use, making it exclusionary

## TRAINING

- Should we train people regarding the "safe" use of social media?

## REPUTATION

- Has a reputation of being "evil" and exploiting vulnerabilities
- Has a reputation of being "hijacked" by certain user groups to exploit others
- Reputation of being misused by the media, for example
- Social Media already has a terrible reputation
- It's used for "good" and "bad" - how do we ethically deal with this?

## TEAMING

- Is a teammate in some situation – such as providing vital information (examples being travel advice, immigration applications)

# YOUR PHONE

## TRANSPARENCY

- Not sure about transparency - does the phone clearly communicate to us all of what information is being collected/used?

## CONTROL

- Where does your phone end and where do the applications begin? Who is responsible?
- Multiple people involved, complicated layers and hierarchies
- Right to Repair laws affect your level of ownership over the phone
- Uncontrolled updates

## EASY-TO-USE INTERFACE

- The interface is easy to use, but not easy to understand when it comes to data storage
- 'Easy to use' is relative - many people don't know how to find 'off' for functions, for example

## COMMUNICATION

- Siri might be considered a form of "rich communication" - she speaks like a human

## TRAINING

- Education required towards understanding the nature of autonomy

## REPUTATION

- The system itself may be considered for its dangers as 'not trusted' but other groups may seem trustworthy

## TEAMING

- Siri is a teammate because it is often regarded as being human-like
- My phone is my teammate for navigation
- Phone as a teammate: using blocking applications to keep us off social media, connect with communities (especially during lockdown)
- "Find my iPhone" – it's a team member in search when the phone is lost
- NHS application is a teammate in protecting family/friends from COVID-19

# EXERCISE THREE





# USERS

- Corporate events or art installations to inform on risks of mismanagement of trust
- Develop and promote non-partisan bodies to create recognisable 'trustworthy' standards
- Interactive advertisements, advertising using locative technology - ARG style
- Viral marketing campaigns
- People tend not to care about things unless they've been explicitly shown how it affects them and their lives
- Create more popular culture around it - Black Mirror made an impact
- A compelling mural or public display
- Examples/case studies to communicate important risks/ considerations
  - Examples of reputational damage - loss of employment, safety, belonging, high stakes scenarios, testimonials from others who've experienced it
  - Present the worst case scenarios, like images on cigarette packets, for example
- Real life demos of how easy it is to corrupt systems
- Interactive role-play activities
- TikTok/YouTube
- Digital chatbots or conversational agents that provide prompts

# DESIGNERS

- Provide an explanation of how their designs/ideas/products will be misused
- “The more thought you put into this now - the less you'll have to worry about it later on”
  - Examples of how it impacts them or their job if things go wrong
- Workshops to help them engage with the customer directly to better understand their needs – interactive engagement sessions with user groups
- Give them an experience of helplessness – put them in users’ shoes
- Escape room where the designers have to face similar issues that the users face
- Case studies
- Explain how designing for trust creates a good user experience
  - Trust means different things to different people in different contexts and needs to be considered as part of design
- Build responsibility for the use of the product into the design brief
  - Make user testing - with a focus on misuse and experimentation - part of the design process
- Make named individuals within organisations responsible a product meeting design criteria.
  - Make named individuals liable for their products

# ENGINEERS

- Give engineers the opportunity to interact with customers/end users to inform their work
- Role playing scenarios where it could go wrong
- Nominate directly responsible individuals within teams, who are contractually liable, and ethically responsible for components
- “Use your power for good”
- “You cannot request trust, you have to earn it”
- Interactive game to communicate concepts/risks
- Problem-solving puzzles
- How do you create a culture and infrastructure around trust?
- Create an AI that's good enough to notice where engineers are building with disregard to trust and then flag it - transparency and social pressure of in-group values
- Need ethics by design - and include requirements in contracts/sanctions for not doing so
- Greater trust = more buy in, more sellable data
  - Lower trust = lower ROI
- Have team building exercises where engineers design untrustworthy models that affect each other - like a Prisoners' Dilemma situation to learn from a contained experience

# CONCLUSION

# POSSIBLE IDEAS TO TAKE FORWARD

- Interactive role playing scenarios where it could go wrong
- Interactive game to communicate concepts/risks
- Problem-solving puzzles
- Have team building exercises where engineers design untrustworthy models that affect each other - like a Prisoners' Dilemma situation to learn from a contained experience
- Provide an explanation of how their designs/ideas/products will be misused
- Workshops to help them engage with the customer directly to better understand their needs – interactive engagement sessions with user groups
- Give them an experience of helplessness – put them in users' shoes
- Escape room where the designers have to face similar issues that the users face
- Corporate events or art installations to inform on risks of mismanagement of trust
- Interactive advertisements, advertising using locative technology - ARG style
- Examples/case studies to communicate important risks/ considerations
  - Examples of reputational damage - loss of employment, safety, belonging, high stakes scenarios, testimonials from others who've experienced it
  - Present the worst case scenarios, like images on cigarette packets, for example
- Digital chatbots or conversational agents that provide prompts

# POSSIBLE FOLLOW-UP ACTIVITIES

- Re-run the workshop with new audiences – Doctoral Training Networks are a good example
- Create a public engagement activity using the ideas mentioned in the previous slide
  - Our connections with the TAS Hub and now, the N-TAIL group, could help with access to newer audiences, as well as support in running a more interactive workshop or creating something public-facing
- The workshop will be showcased in the TAS Hub newsletter - it could also be publicised as by Thales on our website/on the Intranet

THALES