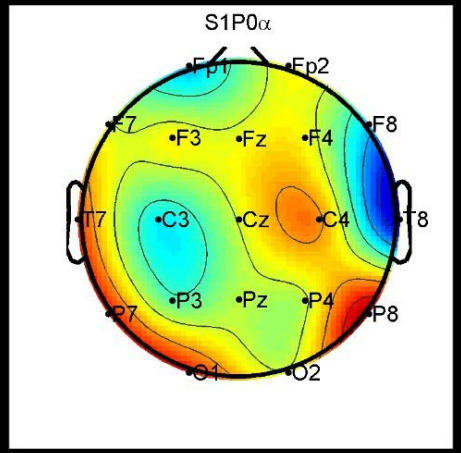
On TRUST in Humans and Machines

Prof. Hussein A. Abbass
University of New South Wales – Canberra

Establishing the Context

Historical Context

Cognitive-Cyber Symbiosis
CoCyS



Cognitive-Cyber Symbiosis
CoCyS

Collect Human Data

Collect Environment Data

Process Signals

Process Signals

Map Features into Mental & Environmental States

Extract Features

Extract Features

Map Features into Environmental States

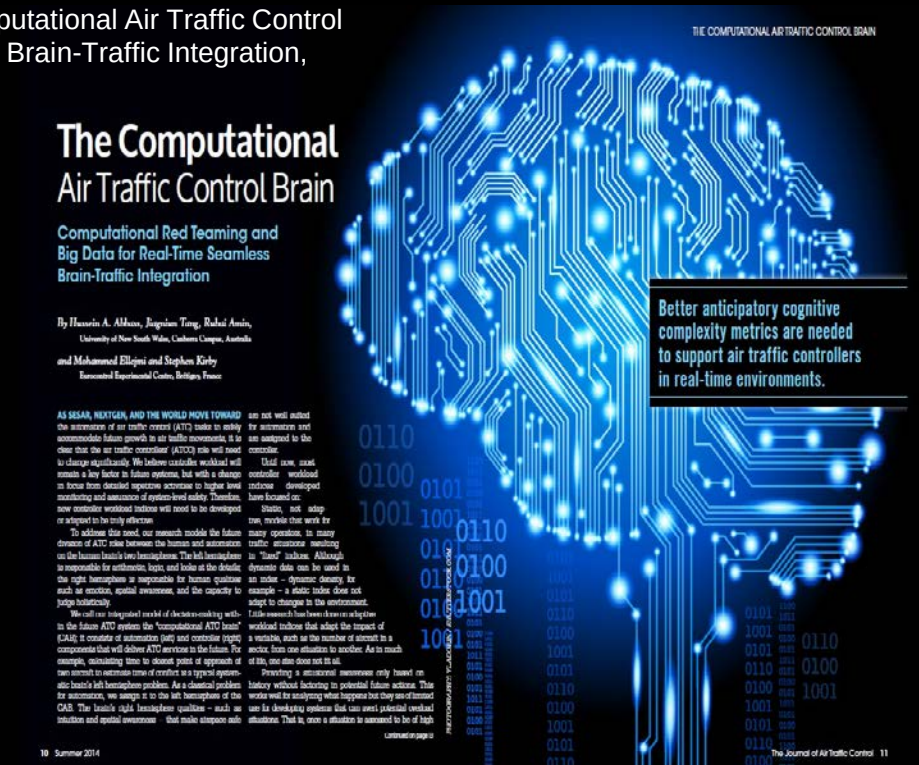
Optimization

Simulation

Red Team

Cognitive-Cyber Symbiosis
CoCyS

Abbass H.A., Tang J., Amin R., Ellejmi M., and Kirby S. (2014). The Computational Air Traffic Control Brain: Computational Red Teaming and Big Data for Real-time Seamless Brain-Traffic Integration, Journal of Air Traffic Control, Summer, 2014.



THE COMPUTATIONAL AIR TRAFFIC CONTROL BRAIN

The Computational Air Traffic Control Brain

Computational Red Teaming and Big Data for Real-Time Seamless Brain-Traffic Integration

By Hussein A. Abbass, Jiaqiao Tang, Rahul Amin, University of New South Wales, Canberra Campus, Australia
and Mohammed Ellejmi and Stephen Kirby, Eurocontrol Experimental Centre, Bretigny, France

AS SESAR, NEXTGEN, AND THE WORLD MOVE TOWARD the automation of air traffic control (ATC) tasks to safely accommodate future growth in air traffic movements, it is clear that the air traffic controllers' (ATCO) role will need to change significantly. We believe controller workload will remain a key factor in future systems, but with a change in focus from detailed supervisory activities to higher level monitoring and assurance of system-level safety. Therefore, new controller workload indices will need to be developed or adapted to be truly effective.

To address this need, our research models the future division of ATC roles between the human and automation on the human brain's two hemispheres. The left hemisphere is responsible for arithmetic, logic, and looks at the details; the right hemisphere is responsible for human qualities such as emotion, spatial awareness, and the capacity to judge holistically.

We call our integrated model of decision-making within the future ATO system the "computational ATO brain" (CAB); it consists of automation (left) and controller (right) components that will deliver ATO services in the future. For example, calculating time to closest point of approach of two aircraft to estimate time of conflict is a typical systematic brain's left hemisphere problem. As a classical problem for automation, we assign it to the left hemisphere of the CAB. The brain's right hemisphere qualities – such as intuition and spatial awareness – that make airspace safe are not well suited for automation and are assigned to the controller.

Until now, most controller workload indices developed have focused on:

Static, not adaptive, models that work for many operations. In many traffic situations resulting in "fixed" indices. Although dynamic data can be used in an index – dynamic density, for example – a static index does not adapt to changes in the environment.

Little research has been done on adaptive workload indices that adapt the impact of a variable, such as the number of aircraft in a sector, from one situation to another. As in much of life, one size does not fit all.

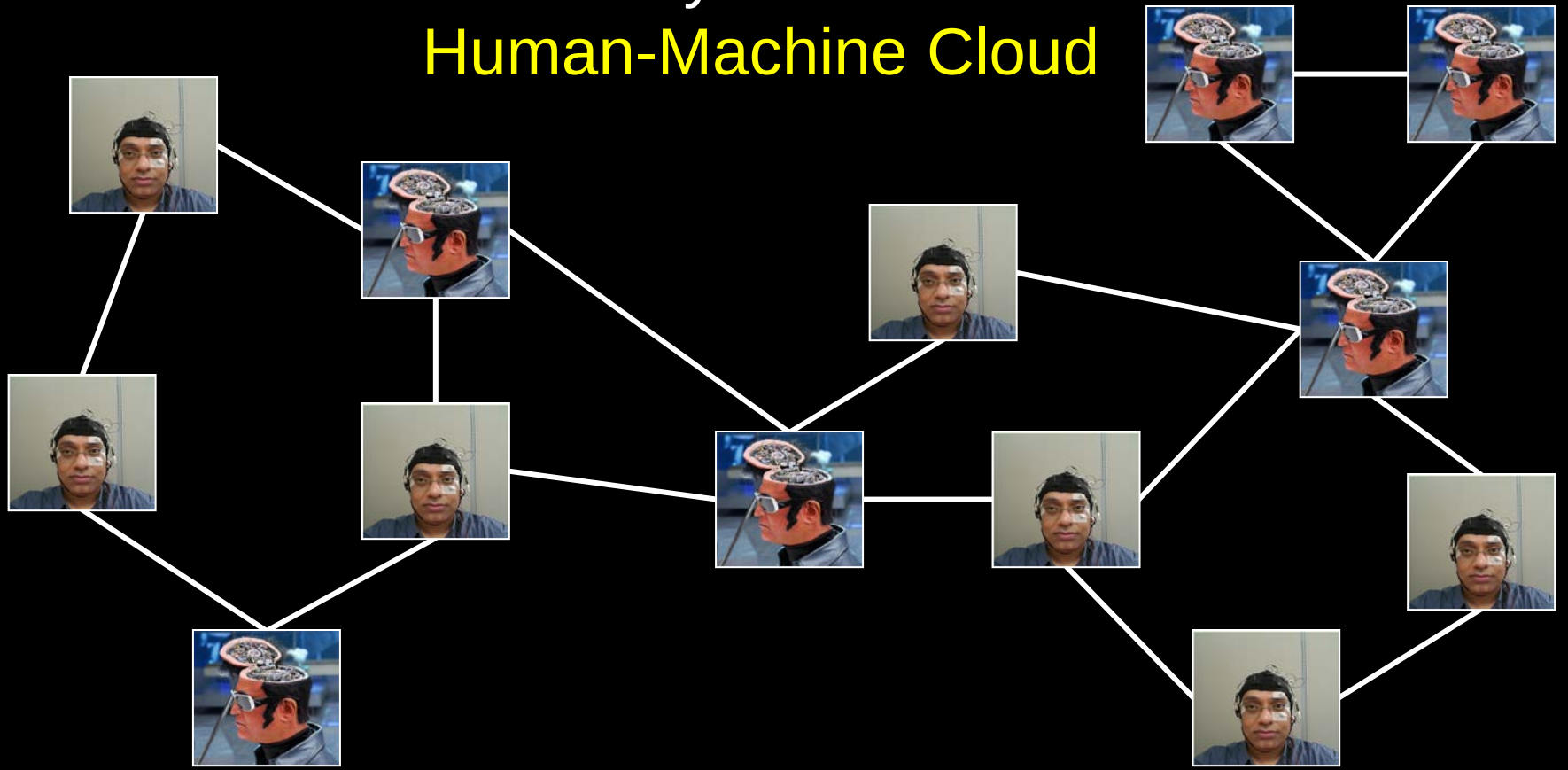
Perceiving a situational awareness only based on history without factoring in potential future actions. This works well for analyzing what happens but they are of limited use for developing systems that can avert potential overload situations. That is, once a situation is assessed to be of high

Continued on page 12

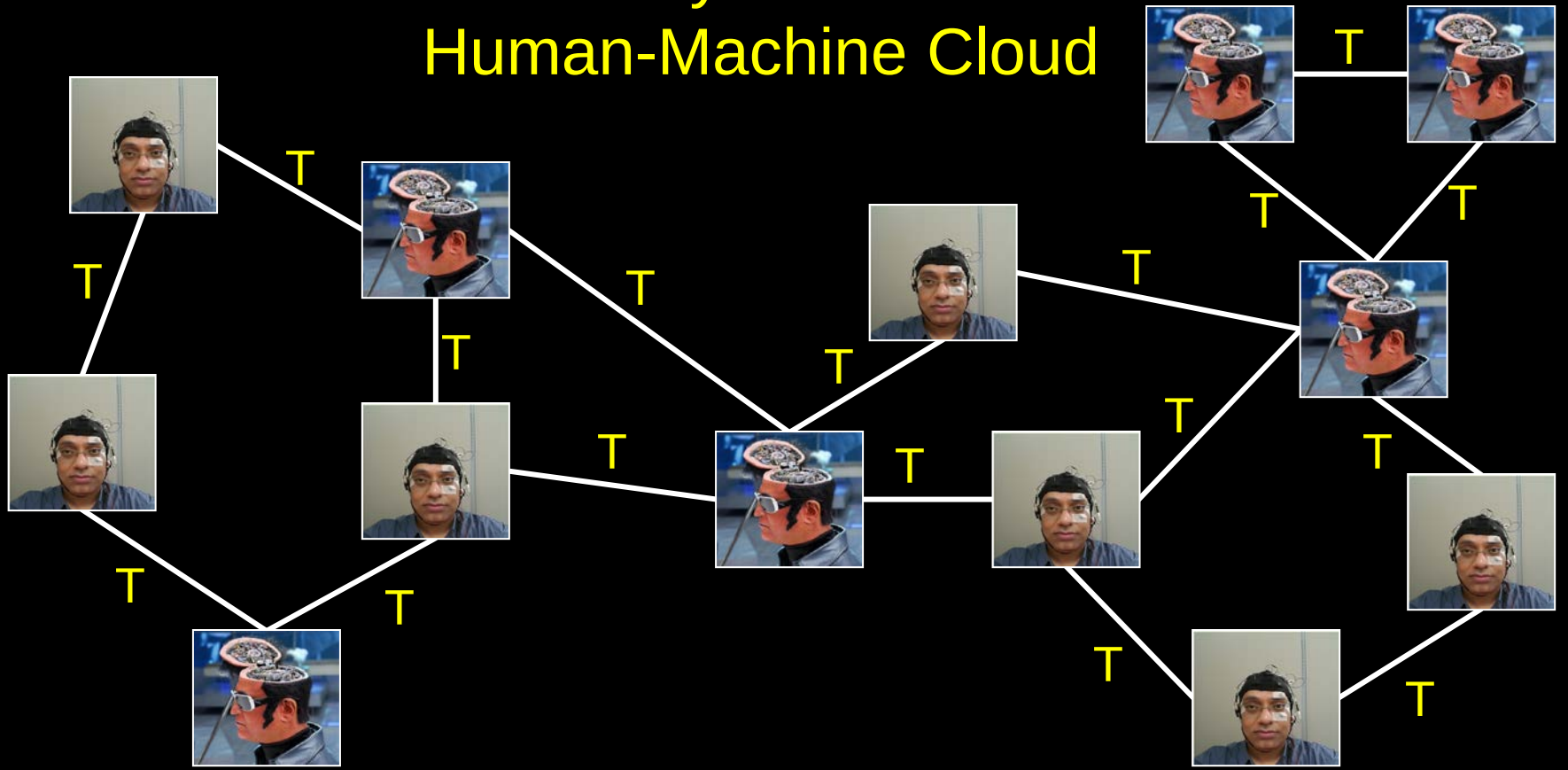
Better anticipatory cognitive complexity metrics are needed to support air traffic controllers in real-time environments.

Vision

CoCyS' Vision
Human-Machine Cloud



CoCyS Vision
Human-Machine Cloud



The human-machine balance

Inventive work from the **University of New South Wales** is redefining human and artificial intelligence in an effort to create next-generation human-machine symbiosis

ALTHOUGH SELF-GOVERNING machines capable of learning are frequently featured in futuristic films, many are not aware of the real abilities of artificial intelligence. Intelligent agents that can learn and make informed decisions are already used in numerous areas, such as traffic control. Indeed, the Institute of Electrical and Electronics Engineers (IEEE) has predicted widespread use of semi-autonomous vehicles within the next quarter-century. As intelligent computer systems play an increasingly pivotal role in the modern world, it is important that the decision-making capabilities of both humans and computers are improved.

At the University of New South Wales (UNSW), Professor Hussein Abbass is endeavouring to understand how models can be developed that allow for improved decision-making processes by redefining intelligence and thus paving the way towards trusted autonomous systems.

DEVIL'S ADVOCATE

After more than a decade of researching the nature of competition and competitive interaction, Abbass' work has led to Computational Red Teaming (CRT), a state-of-the-art architecture to support decision making. The foundational concept of CRT is that human intelligence is derived from the ability to calculate risk and push boundaries by challenging an environment. "To have a true intelligent system, we need mechanisms to assess risks and to design and create challenges," explains Abbass.

> It is hoped that this research, once distilled into computational models, can be embedded within decision-making models to design a trusted system for human-computer interaction

In the strategic concept of Red Teaming (RT), individuals look at their own decisions through the eyes of direct competitors to make strategic assessments. CRT employs multi-agent systems (MAS) and computational intelligence (CI) techniques in an attempt to transform this 'devil's advocate' approach into systemic, computable steps to improve decision-making processes. Whereas a classic computer program will be written to solve a problem, CRT involves programs that are developed with the ability to define their own programs, something Abbass likes to call 'meta-programming'. For example, in order to calculate risk, an agent can write its own objective while programmed with the mechanisms that allow it to define which uncertainties are most relevant and how these might impact the objective.

Implemented properly, CRT can be used to explore uncertainties, locate vulnerabilities, learn about other entities in the environment, understand biases, access information on other relevant decision cases and unlearn in order to learn. It is even able to explore ideas and scenarios that humans would not be able to process in the same timeframe.

As a proof of concept, CRT has been employed with success in an air traffic control (ATC) scenario. Electroencephalography has allowed Abbass to continuously measure and analyse the brain signals of air traffic controllers while simultaneously analysing air traffic information in real time. Using cues from either one or both data sources, a decision can then be made about a course of action, clearly showing the benefits of CRT as a decision support system in the context of an increasingly automated ATC environment.

HUMAN–MACHINE INTERFACE

The international use of CRT speaks volumes about the role of human mental processes within automation. "I have connected the human brain to the complex air-traffic environment so that automation works in harmony with human cognitive abilities," Abbass elaborates. In working towards a next-generation form of intelligence, he envisions a team of humans and one of machines collaborating harmoniously to solve problems and make decisions. This vision is called Cognitive Cyber Symbiosis (CoCyS). Both human and machine 'thinking' are processes carried out within the electromagnetic spectrum, so why not blend cognitive space and cyber space together and transfer thoughts autonomously and seamlessly?

The precursors of what may sound like a fantastical proposition can already be glimpsed in today's world. One only needs to consider the brain-computer interfaces that allow disabled users to control motorised wheelchairs. What Abbass has in mind, however, is more accurately described as a next-generation human–machine cloud. Through the symbiosis of human–computer thinking, CoCyS aims to speed up the communication channel between humans and computers to bring about a superior real-time, evidence-based decision-making process.

Currently, CoCyS is a puzzle for Abbass – wherein all the pieces have been identified but do not yet fit together. Due to the sheer complexity of human brain signalling, responding constantly to innumerable sensory stimuli, a large ambiguity arises in the communication channels. Real-time, *in situ* data cleaning of brain signalling in complex situations is not yet possible, but instead of removing ambiguity, Abbass proposes to manage it. If person A tells person B something but the meaning is unclear, person B can ask questions that increasingly reduce the ambiguity, effectively cleaning up the signal. Interaction, therefore, is the key.

TRUSTED AUTONOMY

Before humans and computers can work together seamlessly, there are some trust issues that need to be addressed. In every domain that relies on reciprocal interaction between agents, trust plays a critical role – and yet a true understanding of the dynamics of trust remains elusive. Equally important to both human–machine and human–human interaction, Abbass' work attempts to understand how trust plays a pivotal role in designing an environment in which the CoCyS dream can become a reality.

To elucidate the dynamics of trust, Abbass wants to understand how it is reinforced in society and how it is transferred through the development of game theoretic models. In classical games, the decision-making process is carried out in a way that denies researchers the chance to study the role of influence, whereas the decision-making process in trust games allows this. There is little research regarding strategies to influence and transfer trust, but CRT can be used to provide important insights. Abbass' main goal is to discover whether a strategy can be employed by a truster to change an unreliable trustee into a reliable one and what those strategies are. It is hoped that this research, once distilled into computational models, can be embedded within decision-making models to design a trusted system for human-computer interaction.

RESHAPING INTELLIGENCE

OBJECTIVES

To redefine human and artificial intelligence and improve decision-making processes through:

- Computational Red Teaming (CRT): analysing challenge and risk in humans and machines
- Cognitive Cyber Symbiosis (CoCyS): connecting humans with cyber space
- Trusted autonomy: investigating the role of trust in a cooperative human–machine environment

KEY COLLABORATORS

Dr Sameer Alam, Dr Michael Barlow, Dr George Leu, Dr Kathryn Merrick, Associate Professor Tahid Sattar, Dr Kamran Shafi, Dr Jiaqiao Tang, University of New South Wales, Australia • **Dr Axel Bender, Dr Svetoslav Goldov,** Defence Science and Technology Organisation, Australia • **Professor Annaswamy Swaminathan,** SINAPSE, National University of Singapore • **Professor David G Green,** Monash University, Australia • **Professor Garrison Greenwood,** Portland State University, USA • **Dr Eleni Petraki,** University of Canberra, Australia • **Associate Professor Kay Chen Tan,** National University of Singapore • **Professor Xin Yao,** University of Birmingham, UK

FUNDING

Australian Research Council

CONTACT

Professor Russell Abbass

University of New South Wales
School of Engineering and Information Technology
Northcott Drive, Canberra
ACT 2600, Australia

T +61 262 688 158
E h.abbass@unsw.edu.au
E hussein.abbass@gmail.com

www.husseinabbass.eu

PROFESSOR HUSSEIN ABBASS has worked for the past 25 years on characterising and connecting human and artificial intelligence. His objective is to design dual-use models that can both be implemented as pencil-and-paper tools as well as sophisticated autonomous smart computer systems for organisations. The main driver for his research is to improve decision-making on all levels, from individuals to governments.

UNSW AUSTRALIA Canberra

On Trust

Petraki E. and Abbass H.A. (2014) On Trust and Influence: A Computational Red Teaming Game Theoretic Perspective. IEEE Computational Intelligence in Defence and Security Symposium, Hanoi, December 2014.

Abbass H.A., Greenwood G., & Petraki E. (conditionally accepted) The N-Player Trust Game and its Replicator Dynamics. IEEE Transactions on Evolutionary Computation.

Abbass H.A., Tang J., & Petraki E. (under review) Shaping Influence and Influencing Shaping: a computational red teaming trust-based model, MODSIM 2015.

Organisational Psychology of Trust

- "Willingness to be vulnerable to another based on the expectation of favorable outcomes for the trusting party"
  - [Mayer, Davis & Schoorman; 1995]
- Trust "introduces unwanted uncertainty into our lives". It means that other people control outcomes that we value. It gives people "power over us"
  - [Kipnis; 1996, p. 40]

Organisational Psychology of Trust

- Trust is having "expectations, assumptions or beliefs about the likelihood that another's future actions would be beneficial, favourable or at least not detrimental to one's interests'
  - [Morrison and Robinson; 1997] (p, 238).
- Distrust deteriorates overall organization performance [Karl; 2000]

Organisational Psychology of Trust

- Whitener, Brodt, Korsgaard, & Werner found that trust in the workplace is linked to team cooperation, performance, and quality of communication in organizations.
- It has been suggested that the more the employees trust their managers, the more satisfied they are with their jobs and this leads to their general good health and wellbeing [Helliwell, Huang & Haifang; 2011]

Cognitive Psychology of Trust

- Kim argues that interpersonal trust is based upon listeners' perceptions of a speaker's expertness, reliability, intentions, activeness, personal attractiveness, and the majority opinion of the listener's associates
  - [Kim; 1967]

Cognitive Psychology of Trust

- Deutsch [Delgado, Frank & Phelps; 2005] sets constraints on trust in situations where a person is faced with.
- A person perceives a situation will lead to two events (a path with ambiguity)
  - one she perceives to have negative valency that is greater than the positive valency she perceives to be associated with the second.
- However, which event will occur is reliant on a second person. If the first person chooses this path, she is said to trust the second; otherwise she distrusts the second person.

Sociology of Trust

- The basis to form healthy relationships [Deutsch; 1996, Luhmann; 1979]
- The backbone that glues a social system and acts as a complexity reduction/management mechanism [Luhmann; 1979].
- Luhmann [1979] sees trust as a facilitator for adaptation to occur in a social system, and therefore trust achieves in a social system the equivalent of adaptation in biological systems [Helliwell, Huang & Haifang; 2011].

Neuroscience of Trust

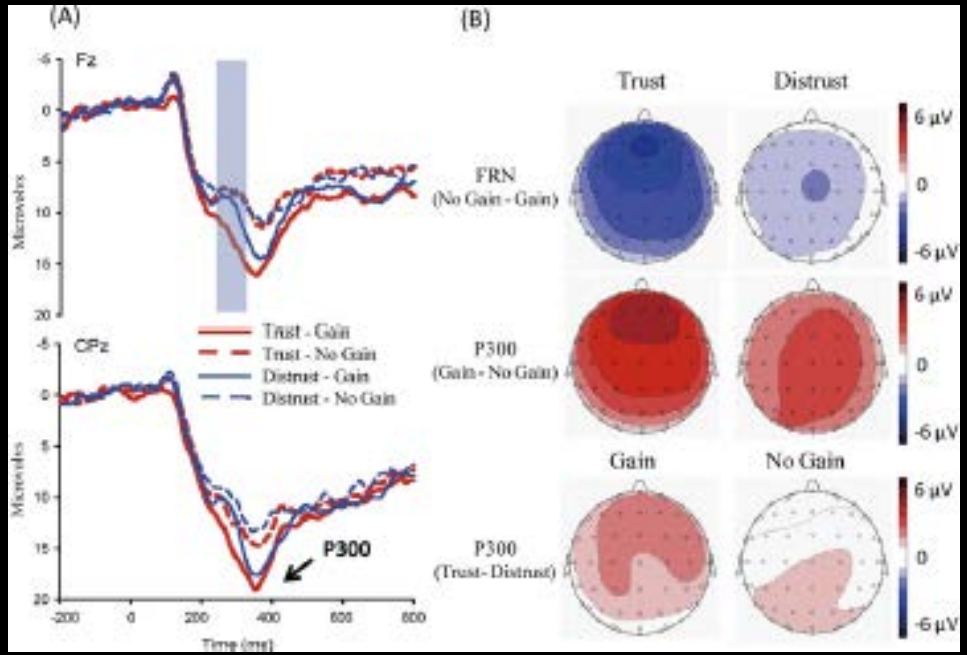
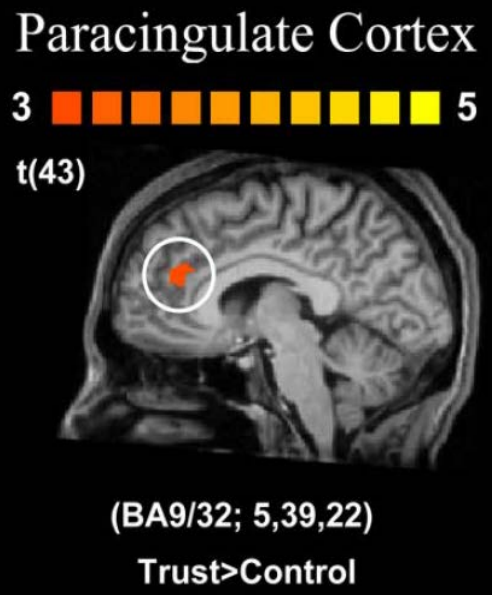


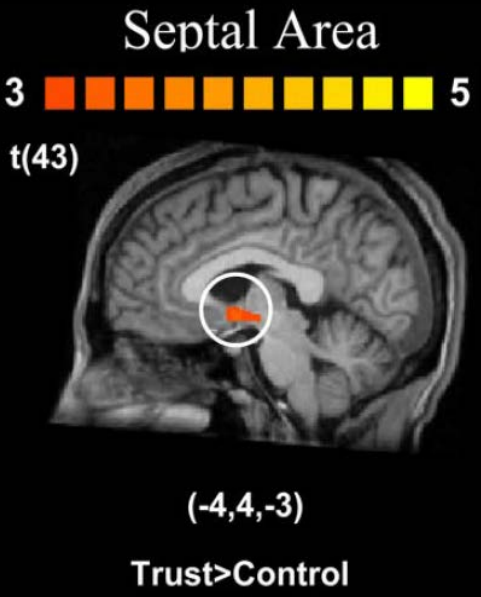
Fig. 4. (A) ERP waveforms time-locked to the onset of feedback stimuli in Experiment 1, sorted according to the participants' choices and outcomes.(B) Scalp topographies of the difference waves between ERP responses to the no-gain vs. gain outcomes averaged for the 230–310 ms time window (the upper panels) and between the gain vs. no-gain outcomes for peak values in the 250–450 ms time window (the lower panels).

Y. Long et al. / Neuroscience 200 (2012) 50–58

Neuroscience of Trust

Thoughts, Feelings, Beliefs





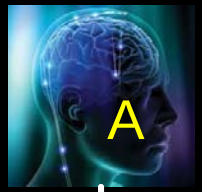
Social memory and learning

Krueger et al. PNAS 104(50),20084-20089, 2007.

Linguistics of Trust

- A perception survey of 115 students by Lam unveiled that “participants trusted leaders who used linguistic politeness strategies in their emails, as opposed to those who failed to include mitigating strategies” [Lam; 2011]
- strategies on the influential nature of trust have revealed that issuing of superfluous apologies can be effective in promoting people sense of trust towards the apologiser [Brooks, Dai & Schweitzer; 2014].

Game Theory of Trust



Trust

No Trust

A ← $10
B ← $0

Untrustworthy

Trustworthy

A ← $0
B ← $100

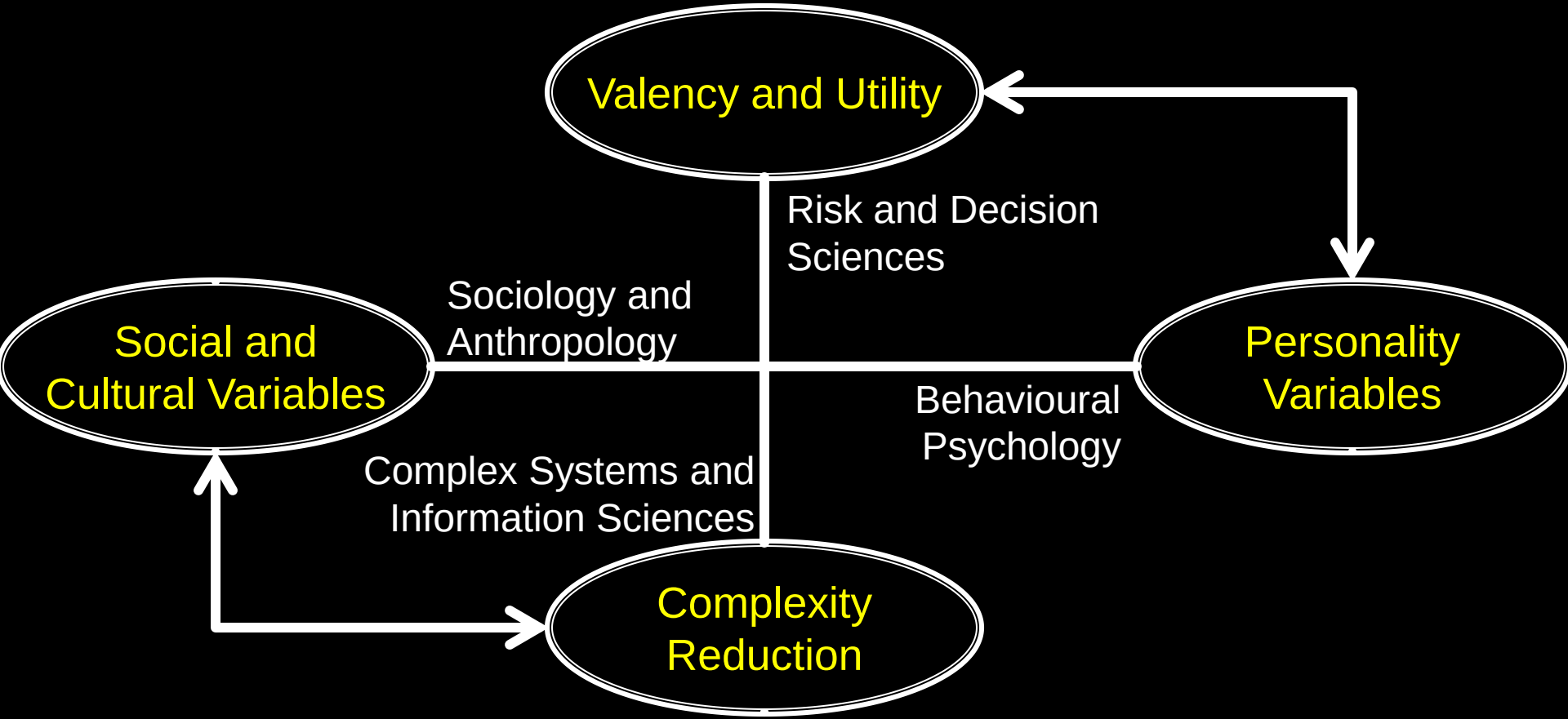
Temptation for Untrustworthiness

A ← $70
B ← $70

Temptation for Trusting

Putting it Together

Putting it Together



Putting it Together

From Humans to Machines

Is a Machine Different?

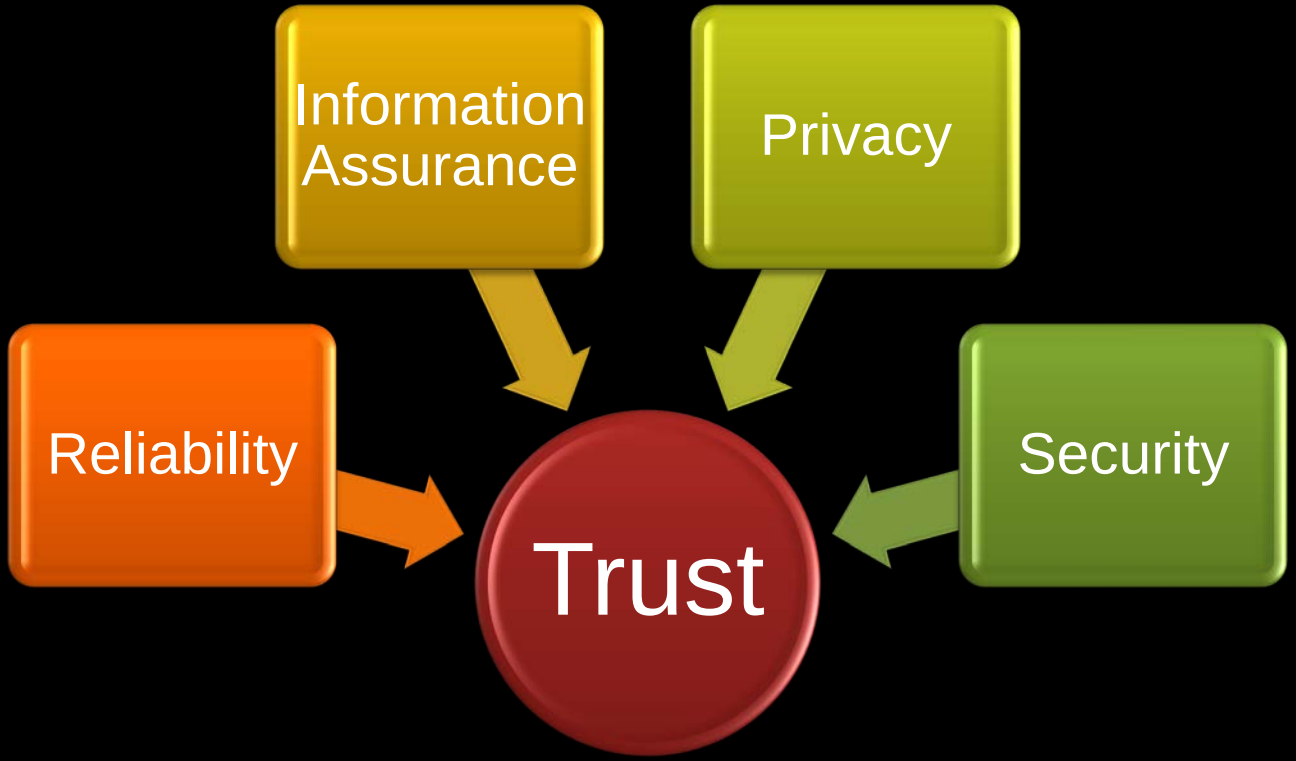
Human

- Conceptual models
- Quantitative models
- Objective/subjective assessment

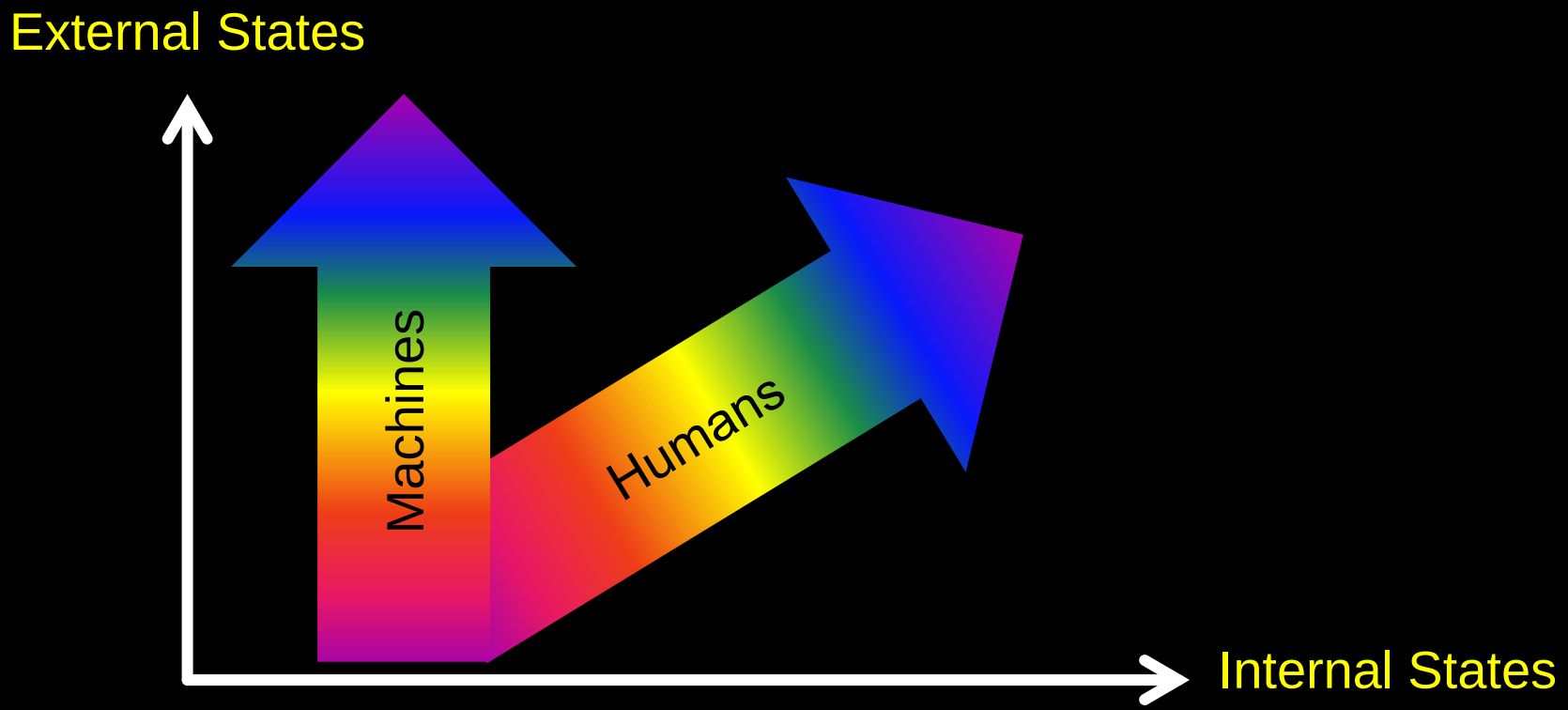
Machine

- Objective quantitative models

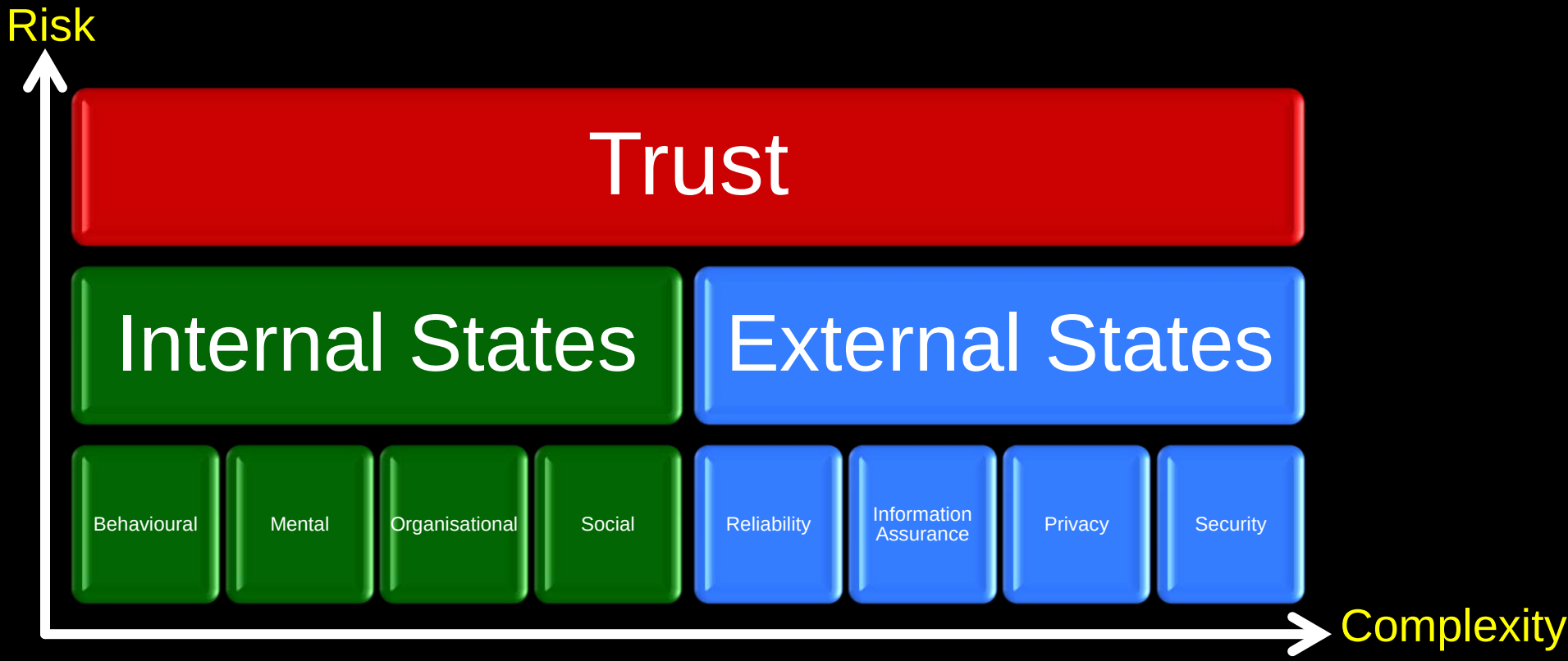
Machine View of Trust



Humans and Machines



Machine View of Trust



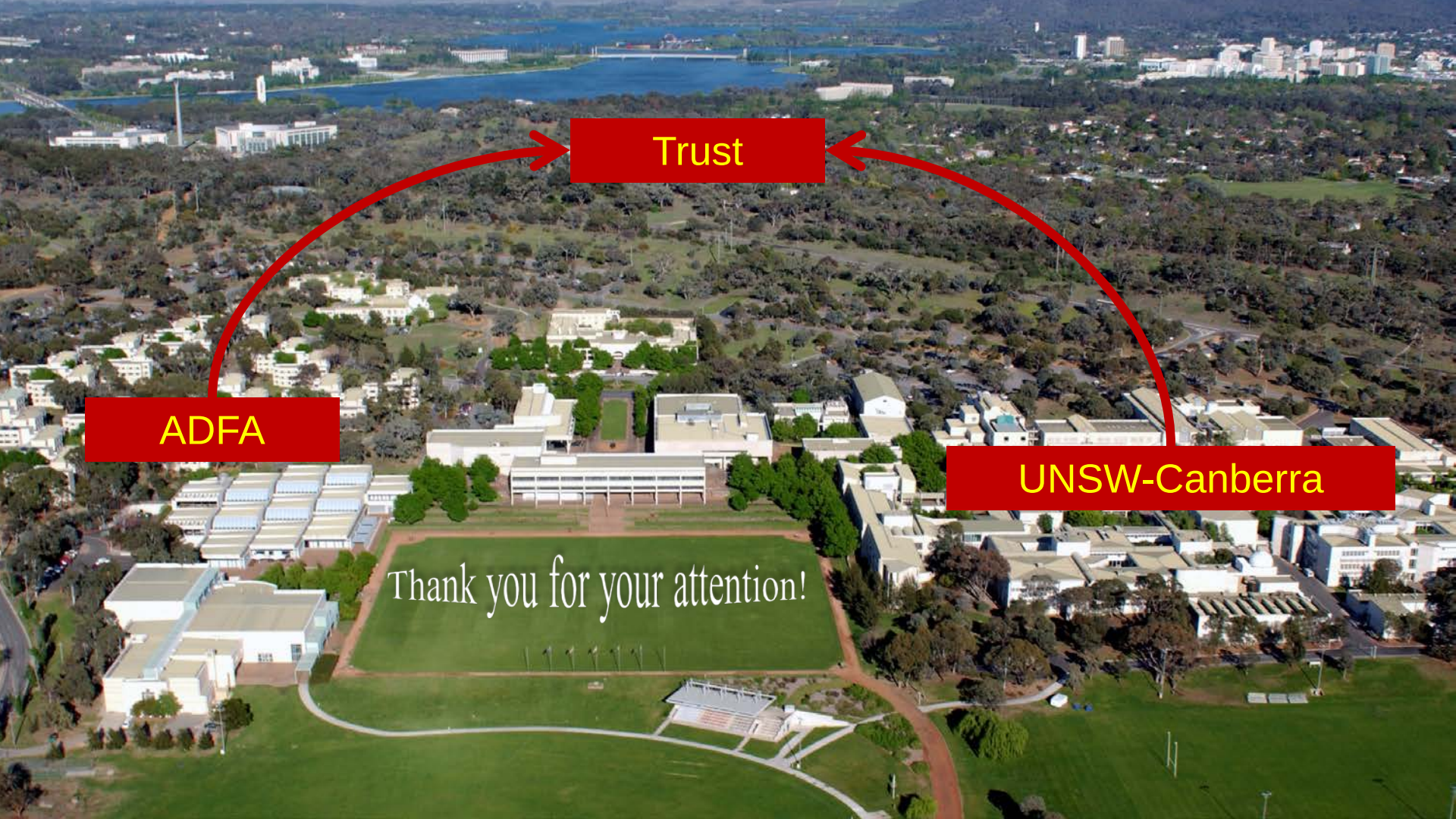
Research Gaps and Opportunities

Sample Research Gaps on Trust

- Western Vs non-Western Data on Trust
- Trust in a multi-cultural setting
- Trust and distrust taxonomies
- Authentic linguistic data, and the analysis of linguistic strategies in naturally occurring interactions that evoke trust
- the lack of linguistic and computational environments to support studies on trust
- Research on trust in neuroscience is nowhere as mature as these classical fields

Sample Research Gaps on Trust

- Indicators on the presence or absence of trust
- Indicators-to-Causes mapping (Atlas of Causes of Trust)
- Prediction of trust or mistrust in a context
  - Maximum look-ahead time for an accurate prediction
  - Continuity of trust
  - Trust as a socially stable phenomenon
  - Context-Prediction-Accuracy relationship
- Shaping causes to influence effects


Trust
ADFA
UNSW-Canberra
Thank you for your attention!