Barriers and Enablers to Measuring Human Trust within AI, Robotic and Autonomous Cyber-Physical Systems

Prof Phil Morgan: Human Factors & Cognitive Science

School of Psychology, Cardiff University, UK

School of Psychology

Director – Human Factors Excellence Research Group (HuFEx)
Director of Research – Centre for AI, Robotics, & Human-Machine Systems (IROHMS)
Human Factors & Cognitive Science + Transportation Lead: CU Digital Transformation Innovation Institute (DTII)
Director – Airbus Centre of Excellence in Human-Centric Cyber Security & Co-Director (H2CS) – Airbus & Cardiff University Partnership
Guest Professor – Luleå University of Technology

HuFEx

Human Factors Excellence Research Group

Cardiff University School of Psychology

Psychology: ~Largest & best resourced in UK

- RAE/ REF (Research Excellence) top 10 since 2001
- >120 Academic, research & prof support staff
- BSc ~950, MSc ~150, PhD >120
- External funding (2014-2023): ≈ >£140m

CORE AREAS:

- Neuroscience (including £68M CUBRIC)
- Cognitive Science & ***Human Factors (since 1965)***
- Developmental & Health Psychology
- Social & Environmental Science

School of Psychology

HuFEx

Augmented & Assistive Systems
Cyberpsychology
Defence & Security
Emergency Services & Healthcare
Humans in AI & Automation
Transportation Human Factors

14 staff (HF, cog sci, social cog, neuroscience)
12 PhD students (AI, automation, cyber security, emergency services, HRI, transport)

~£15m (30+ grants 2017+); ~£5m under review; Phil Morgan ~£37m (50+ grants)

IROHMS

Human-like AI

- Affective computing
- Augmented cognition
- Computational semantics
- Contextual reasoning

Ethical and Explainable AI

- Ethical AI
- Explainable AI
- Explainable robotics
- Trusted autonomy

Human-centred Technologies and Society

- Human-centred computing
- Human-centred cyber security
- Emerging technology and society

Humans and Robots

- Human-centred robotics
- Social robotics
- Robot perception/learning

https://www.cardiff.ac.uk/artificial-intelligence-robotics-and-human-machine-systems

AI for Collective Intelligence (AI4CI)
EPSRC (UKRI): 2024-28

School of Psychology

Director: **Seth Bullock**
Deputy Director: **Nirav Ajmeri**
ECR Lead: **Cangxiong Chen**

A. Heppenstall (Glasgow)
Adam Dennett (UCL)
AI for Smart City Design
With: Arup, UK Govt. …

Leon Danon (Bristol)
AI for Pandemic Resilience
With: Pfizer, DHSC, NHS, Sanofi, Cabinet Office

Academic Pool: 50+ AI/ML co-supervisors from across six AI4CI partner institutions.

Laura Smith (Bath)
Phil Morgan (Cardiff)
Human-Centred Design

Michaela Black (Ulster)
Infrastructure & Governance

AI4CI CORE
Community & Capacity
Cross-Cutting AI/ML
EDI Team
Engagement Team

Hywel Williams (Exeter)
Environmental Intelligence
With: Wikipedia, PML, Met Office, IBM

External Advisory Board
Science Council

John Cartlidge (Bristol)
Maggie Chen (Cardiff)
AI for Financial Stability
With: HSBC, Coinbase…

Aisling O'Kane (Bristol)
Theresa Smith (Bath)
Joan Condell (Ulster)
AI for Health Ecosystems
With: NHS, Mayden

AI4CI Hub

PhD Studentship: Learning to Trust Emerging Disruptive AI and Automated Technology

Cardiff University - Psychology

Barriers & Enablers: Examples

ACCEPTANCE

ADOPTION

CONTINUED OPTIMAL USE
NOT MISUSE, ABUSE ETC.

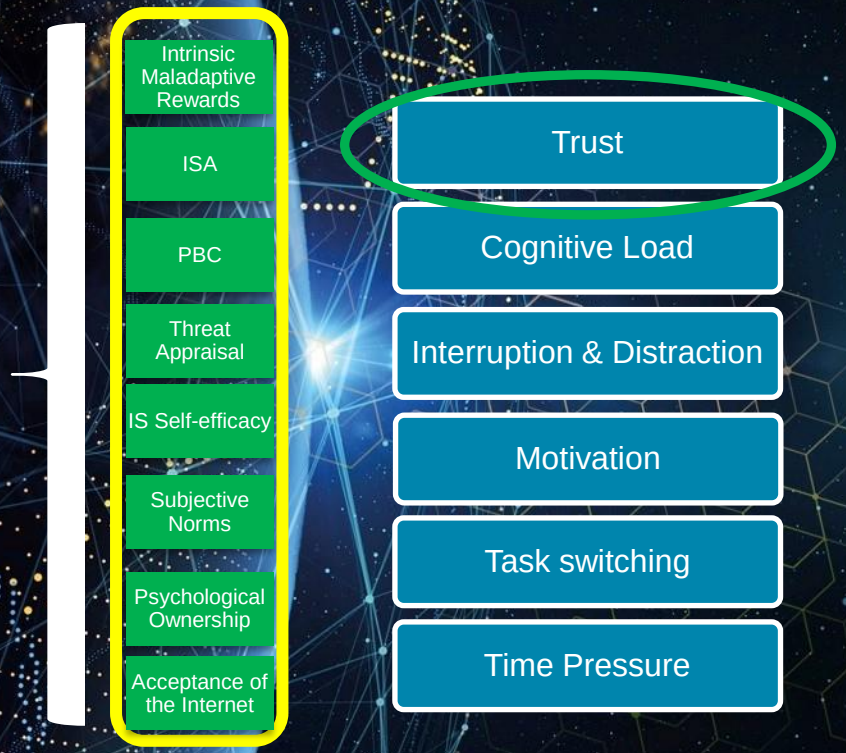
INCIDENTS / ACCIDENTS
TRUST (LOSS, RESTORATION)
RELIABILITY
EXPERIENCE
CHANGE
CYBER SECURITY
PRIVACY
SAFETY
ETHICS
BLAME
RESPONSIBILITY
STANDARDS
CERTIFICATION
LAW & REGULATION
CULTURE
ATTITUDES

AWARENESS
TRAINING
BUY-IN (SELF, COMPANY, GROUP/DIVISION)
COST (£) / COST (OTHER)
FUNDING
MOTIVATION
SELF-EFFICACY
SUBJECTIVE NORMS
PSYCH OWNERSHIP
(MIS)UNDERSTANDING
PRESS & MEDIA
LANGUAGE & COMMUNICATION
MISINFORMATION
RISK(S)

WORKLOAD
TIME PRESSURE
WORKFORCE
WORKLOAD
SITUATION AWARENESS
TASK ALLOCATION
TEAMWORK
ACCESSIBILITY
USABILITY
FUNCTIONALITY
ADAPTABILITY
ASSISTIVE
PRAISE
FEAR
AND MORE…

Example: Developing Metrics & Personas for Optimal Human-Centric Cyber Security

Human Cyber-Security Risk Tool



60%+ Explained

25%+ Explained

Airbus Partnership & Centre of Excellence in Human Centric Cyber Security

AIRBUS

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

School of Psychology

AIRBUS

Transportation Human Factors: Examples

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

School of Psychology

2015-19: **Venturer AVs for UK Roads + Flourish CAVs: Trusted, Secure Mobility.** IUK. With e.g. UoB,FirstBus, UWE, BRL, Atkins, Airbus, AXA, Dynniq, Aiseedo, Designability, Traverse, TSC, AgeUK, Williams (£10.5M)

2019-22: **Decarbonising Transport - Electrification.** EPSRC (~£1M)

2018-19: **Security Perceptions, Emergency Services C2C.** SOS Alarm, Sweden

2020-24: **IDTH Sustainable Transport & IDTH Cyber Security Analytics.** EPSRC, Industry

2019-22: **Centre for AI, Robotics, & Human Machine Systems (IROHMS).** ERDF (>£5M)

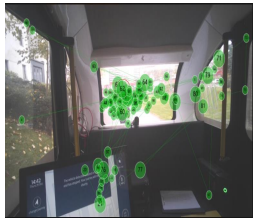
2020-23: **Rule of Law in the Age of AI – UK and Japan:** ESRC JST (~£800k) with Kyoto, Osaka, Doshisha

2023: **Multi-Modal Distraction & Deception.** R-Cloud (with Ksharp)










AVSIMULATION

SCANeR™ Studio

ST Engineering

TRUST
ADOPTION
ACCEPTANCE
CYBER SECURITY
PRIVACY
SAFETY
ETHICS
BLAME
RESPONSIBILITY
LAW & REGULATION

Llywodraeth Cymru Welsh Government
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
JST Japan Science and Technology Agency
EPSRC Engineering and Physical Sciences Research Council
Government Office for Science
SOS Alarm
Innovate UK

Venturer: AVs for UK Roads

VENTURER: ~£5M IUK, 2015-18

HF: Performance, behaviour, individual differences, Situation Awareness, workload, trust, cyber – etc.

Handover of control in **urban settings** = key gap (e.g. Morgan, Alford, & Parkhurst, 2016)

Also: L3-4 (SAE): negotiating traffic, pedestrians, cyclists, responses to AV decision making…

Impact: AV design principles & standards (***safety***), insurance (***legislation, policy***), mobility (***services***), economy (e.g., 10-15k UK jobs - ***employment***).

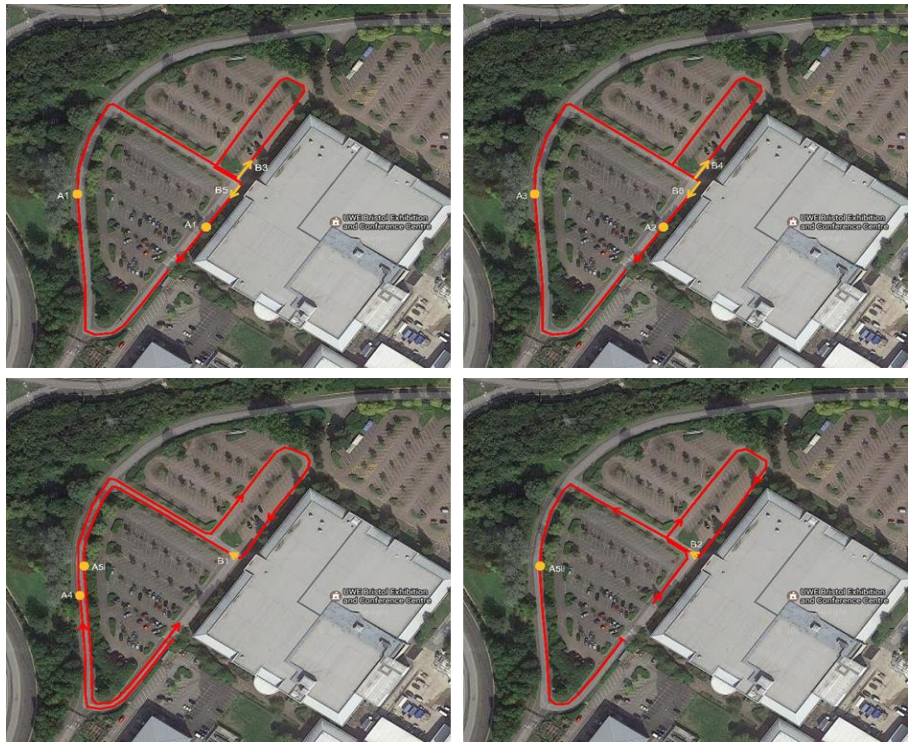
School of Psychology

Humans & AVs - Trust

Trust based on capability? And caution?? Though subjective…

Scenario Type	Scenario Number	Scenario Description	Scenario Picture
A	1	Moving along an empty road at or below the speed limit.	
	3	Overtaking a parked car while leaving a safe distance margin.	
	4	Overtaking a parked car leaving a safe distance margin and waiting if necessary to leave a safe gap from an oncoming car.	
B	1	Turning right off the main road into the side road at a priority junction with no other vehicles.	
	2	Turning right off the main road into the side road at a priority junction with an on-coming vehicle approaching on the main road.	
	3	Turning left out of a side road onto a main road at a priority junction with no other vehicles.	
	4	Turning left out of a side road onto a main road at a priority junction with a vehicle approaching on the main road from the right.	
	5	Turning right out of a side road onto a main road at a priority junction with no other vehicles.	
	6	Turning right out of a side road onto a main road at a priority junction with vehicles approaching along the main road from both directions.	
	7	Turning in left into a side road from a main road at a priority junction with no other vehicles.	

Trial 2: Trust very high (slightly higher within the simulator). Higher during complex & risky maneuvers…!?



Trial 3. Cyclists, pedestrians & vehicle users; higher trust if AV gives way & cautionary.

Parkin, J., Crawford, F., Flower, J., Alford, C., Morgan, P. and Parkhurst, G. (2022). Cyclist and pedestrian trust in automated vehicles: an on-road and simulator trial. *International Journal of Sustainable Transportation*.

Level 4+/5 AVs: Flourish

FLOURISH: Innovate UK, £5.5M, 2016-19

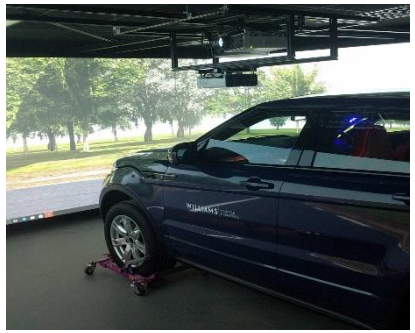
Aims: CAVs & HMIs for those with highest mobility needs (older adults, mobility impaired) incl. as a service

Psych & HF areas: Simulation, usability & UX, trust, workload, SA, HMI design & HCI, cyber security, privacy…

Test **interface interaction** & **responses** incl. eye + HSM (with Airbus): *Stephenson, Eimontaite, Morgan et al. (2021) – Frontiers in Psychology: Performance Science; Voinescu, Morgan et al. (2020). Transportation Research: Part F.*

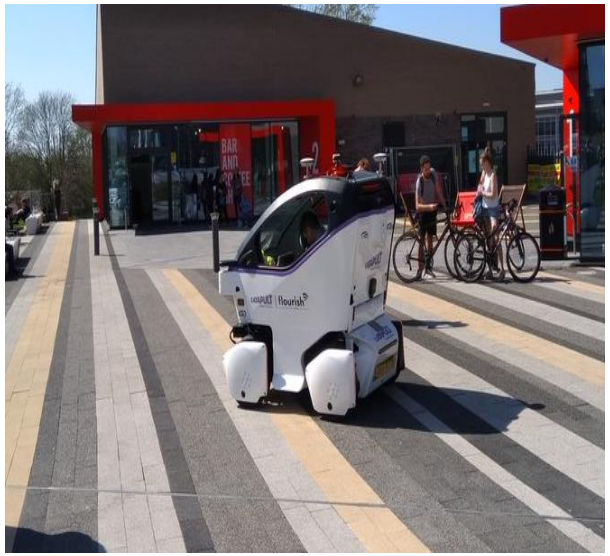
Impact: CAV interface design principles/standards (***safety***), insurance (***legislation, policy***), mobility (***services***), economy (e.g., 6-10k UK jobs - ***employment***), transport as a service (***+++)***

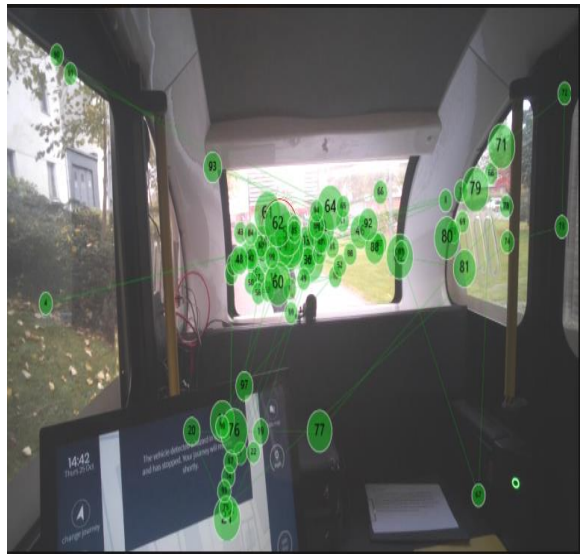
Simulator, Pods, HMI





Design, testing, development & deployment of accessible, usable, functional, adaptable, safe, secure, and trusted human-machine interfaces for connected autonomous vehicles




Trust in CAVs ←→ Trust in CAV HMIs

Voinescu, A., Morgan, P. L., Alford, C., & Caleb-Solly, P. (2020). The utility of psychological measures in evaluating perceived usability of automated vehicle interfaces – a study with older adults. *TR-F: Traffic Psychology & Behaviour 72.*

General trust in tech correlated with HMI usability (key variables in acceptance & attitudes towards AVs (e.g. Liu et al., 2019, Zhang et al., 2019)) & predicts intended AV use (Buckley et al. 2018).

BUT – no relationship with trust in the CAV / Simulator

Challenge: More experience needed (Ekman et al., 2016). Particularly for tech(s) yet to be experienced widely. Ensuring adequate user experience through learning pivotal for success.

PhD Studentship: Learning to Trust Emerging Disruptive AI and Automated Technology (Cardiff University – Psychology)

SIMPLICITY
MINIMAL CLUTTER
ADAPTABLE
ADAPTIVE
ROUTE ADVICE / UPDATES
SYSTEM STATUS
SPEED
SAFE STOP
SERVICE/HELP WHEN NEEDED
EXPLAINABLE & UNDERSTANDABLE

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

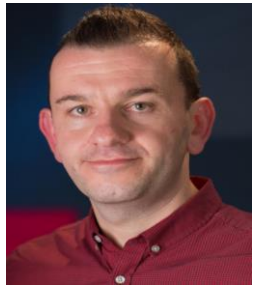
School of Psychology

Rule of Law in the Age of AI: Principles of Distributive Liability for Multi-Agent Societies

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

School of Psychology

TEAM UK (with Profs Bill Macken (2020) & Dylan M Jones OBE (2022))



Prof Phillip Morga

Dr Qiyuan Zhang

Victoria Marcinkiewicz

Louise Bowen

Theo Kozlowski

TEAM JAPAN

Prof Tatsuhiko Inatani

KYOTO UNIVERSITY FOUNDED 1897 KYOTO JAPAN

Prof Minoru Asada

大阪大学 OSAKA UNIVERSITY

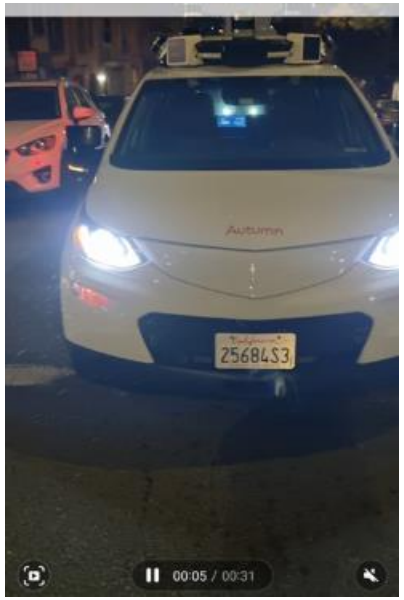
Dr Hirofumi Katsuno

DOSHISHA UNIVERSITY MDCCCLXXV KYOTO VERITAS LIBERABIT VOS

UKRI Economic and Social Research Council

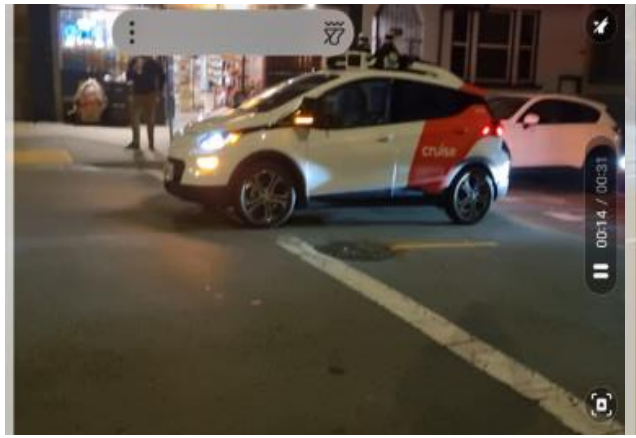
JST Japan Science and Technology Agency

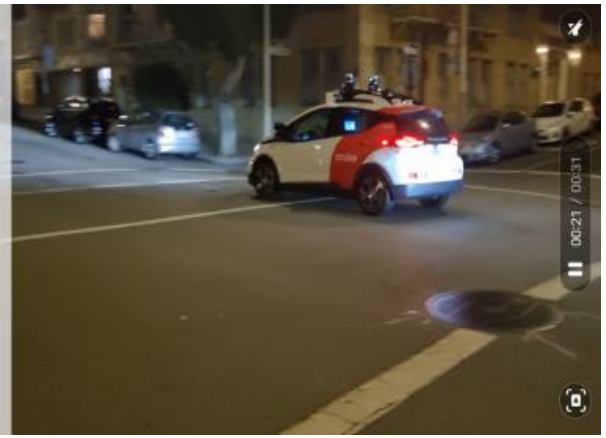
Self-Driving Cars are here…but…



← San Francisco, July 2023 (Morgan, Marcinkiewicz et al.)

Waymo's driverless cars were involved in two crashes and 18 'minor contact events' over 1 million miles





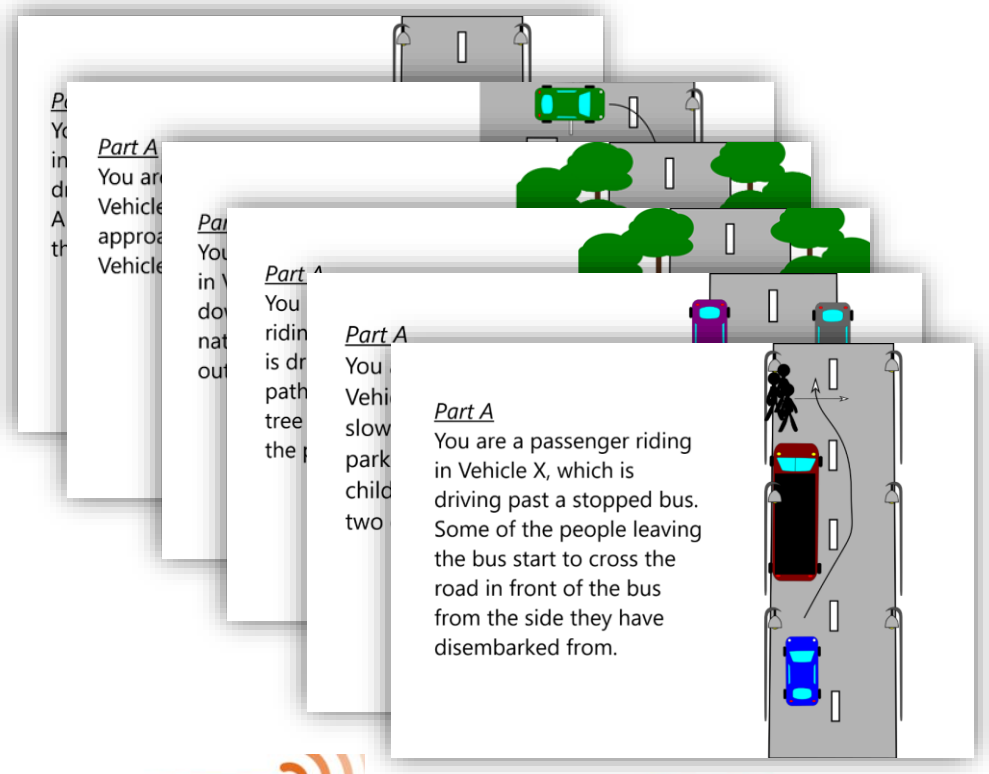
GM's Cruise slashed fleet of robotaxis by 50% in San Francisco after collisions

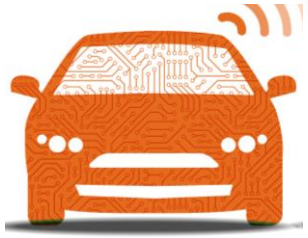
By Samantha Delouya, CNN

Updated 8:01 PM EDT, Tue August 22, 2023

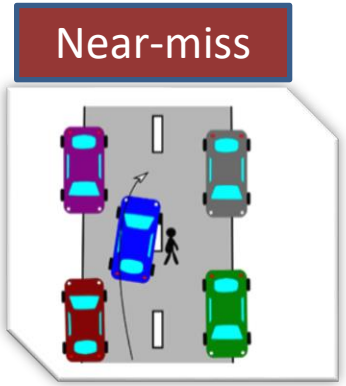
CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD
School of Psychology

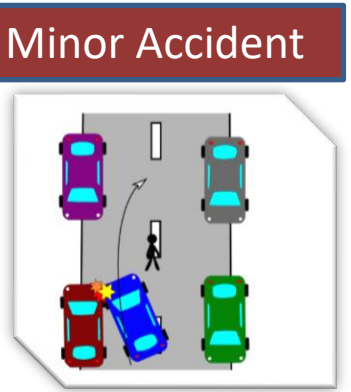
Trust & Blame Before & After Incident

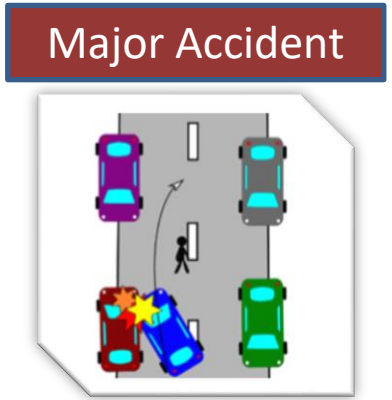




VS







KEY TAKEWAYS (2020-21/22)

- UK: AV blamed more & trusted less;
- Japan: Similar pattern but lower trust;
- Double standards: trust human driver more BUT blame higher vs AV if perceived to be taking a chance;
- 'Ironies': perceived tech & performance capabilities

Zhang, Q., Wallbridge, C. D., Jones, D. M., & Morgan, P. (2021). The blame game: Double standards apply to autonomous vehicle accidents. *Lecture Notes in Networks and Systems*, *270*, 308–314.

Zhang, Q., Wallbridge, C. D., Jones, D. M., & Morgan, P. (2024). Public perception of autonomous vehicle capability determines judgment of blame and trust in road traffic accidents

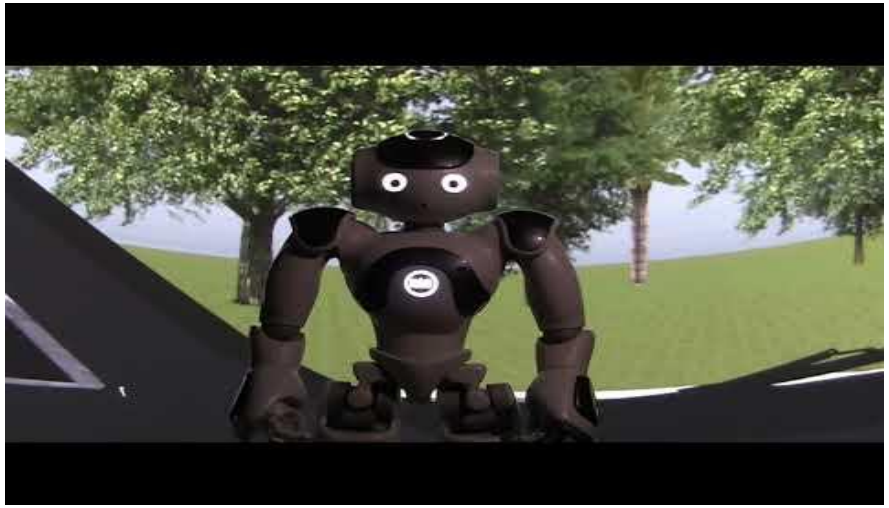
Zhang, Q., Wallbridge, C. D., Jones, D. M., & Morgan, P. (under review). Autonomous vehicle judged less risky and blameworthy relative to a human driver if driven assertively before an accident. *Transportation Research Part A: Policy & Practice*.

Economic and Social Research Council

Japan Science and Technology Agency

Anthropomorphism: Informational Assistants

Highly beneficial for trustworthiness when system(s) running flawlessly – although in incident / accident situations, trust can be damaged more due to the presence of a robot informational assistant



CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD
School of Psychology

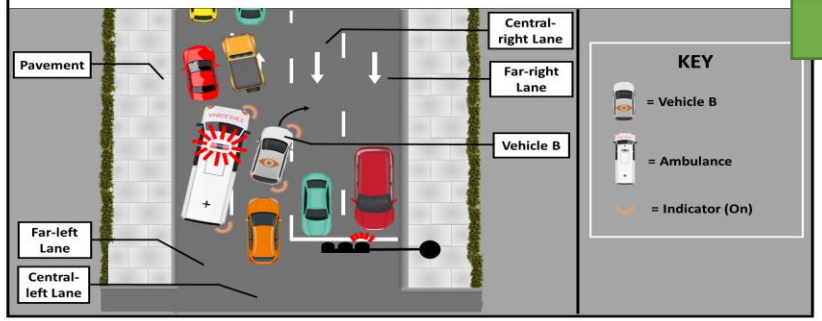
Pushing Boundaries

- SDCs cannot always stop (e.g. emergency situations, environmental factors) & may have / be expected to perform courteous actions;
- The technology is becoming capable;
- But – there will sometimes be negative outcomes.

Part 2

Vehicle B does not stay in the central-left lane. It crosses the broken white line (in the middle of the road) into the central-right lane to give way to the ambulance.

According to the Highway Code, a broken white line marks the centre of the road. A vehicle can cross it if the driver can see the road is clear and wishes to overtake or turn off.

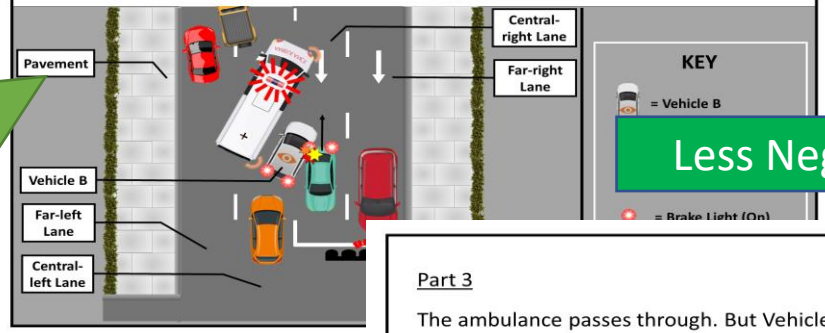


Part 3

The ambulance passes through. But Vehicle B is hit by a vehicle in the central-right lane, which is rolling back because its driver failed to apply the handbrake while waiting at the traffic light.

It is later revealed that if Vehicle B had stayed in the central-left lane, the ambulance would not have been stuck for a long time, and would not have been significantly delayed in its arrival at its destination.

Negative (Accident)

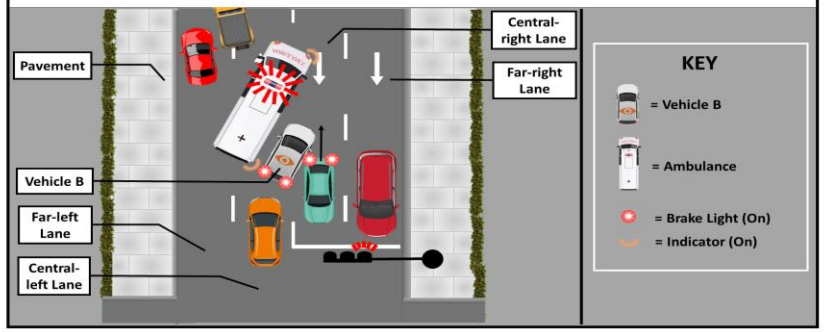


Less Negative (Near-miss)

Part 3

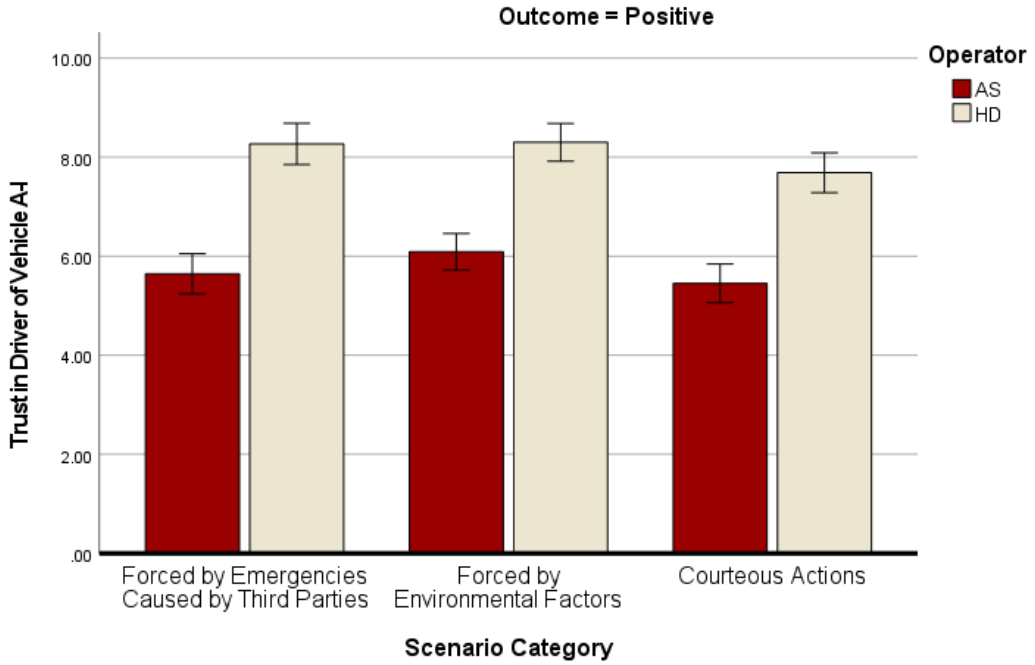
The ambulance passes through. But Vehicle B is nearly hit by a vehicle in the central-right lane, which is rolling back because its driver failed to apply the handbrake while waiting at the traffic light.

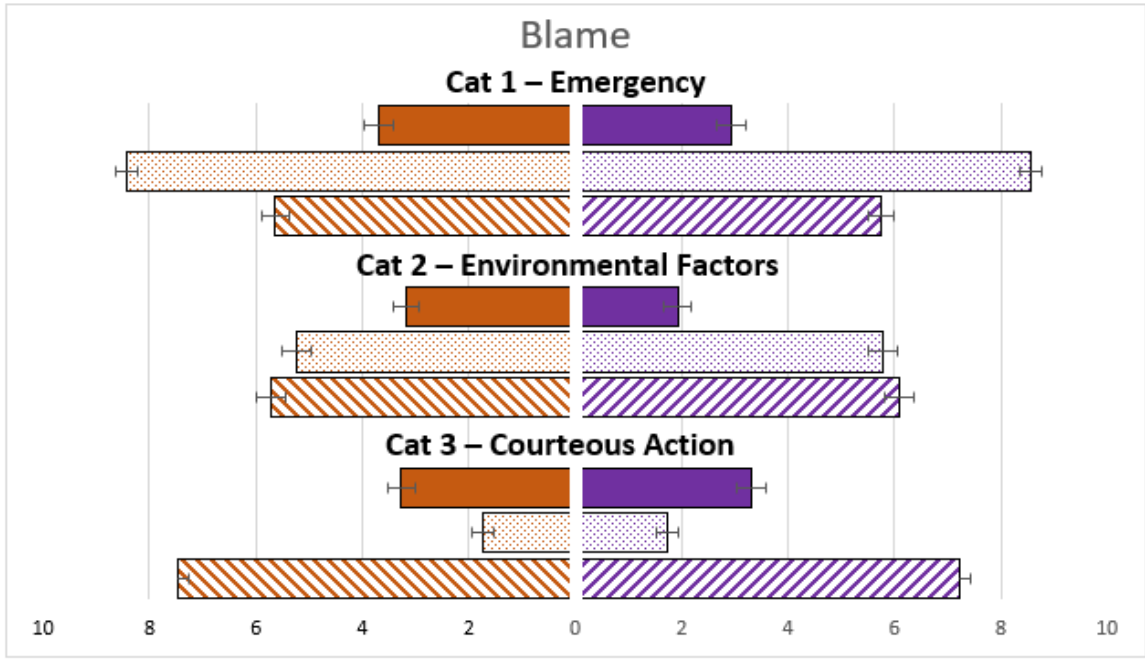
It is later revealed that if Vehicle B had stayed in the central-left lane, the ambulance would have been stuck for a long time, and would have been significantly delayed in its arrival at its destination.



School of Psychology

Trust in Target Vehicle & Blame





Trust higher for HD than AS, $p < .001$

Trust higher for positive outcome than negative outcome, $p < .001$

AND: Trust in AVs increased post vs pre- experiment, especially with positive outcome

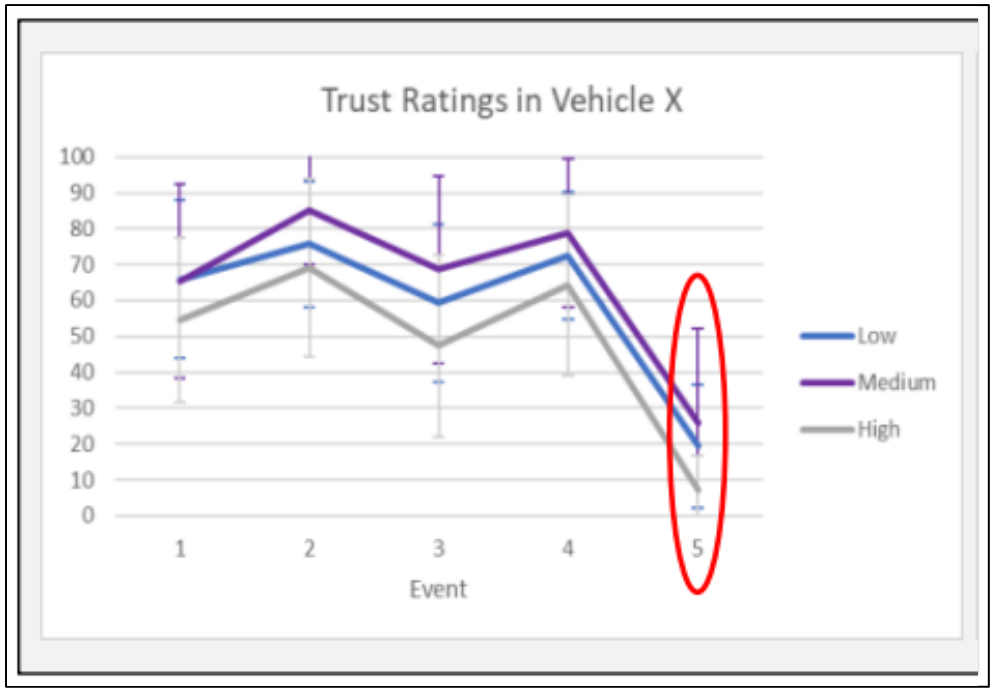
Cyber Security Aspects

Upfront trust in a CAV: impacted by CAV cyber security rating

Upfront trust in the CAV company: impacted by CAV cyber security rating

Trust in CAV and the company after a cyber-attack: *Plummets…can it be restored?*

Trust in CAV after a +/- response to a cyber attack: *Matters…but is it enough?*



School of Psychology

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Defence & Security: Recent Projects

2022-2023: **Measuring Trust in Complex Sociotechnical Systems** HSSRC – with Trimetis

2022-2023: **Developing HF Guidelines for Robots & Autonomous Systems** HSSRC – with QinetiQ & BAE Systems

2023-25: **Multi-Modal Interruption & Distractions** R-Cloud – with K-Sharp

School of Psychology

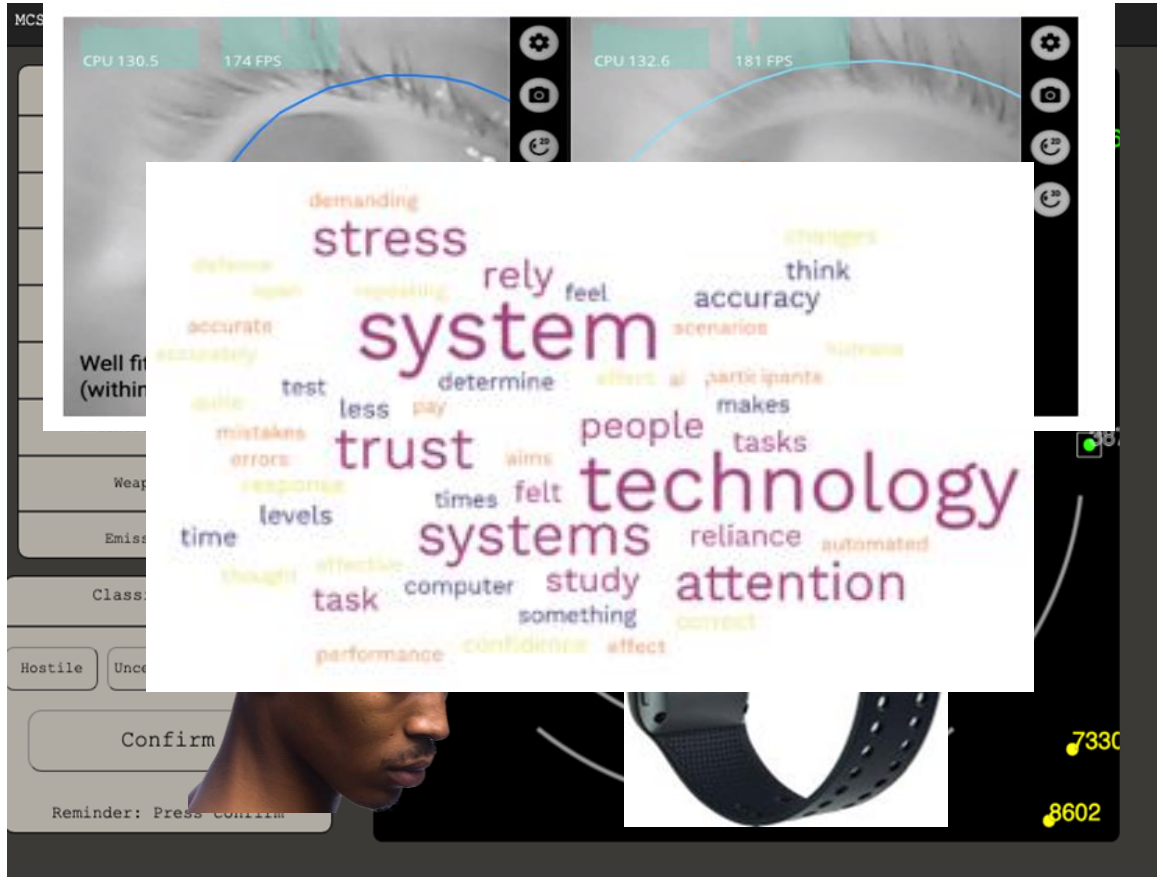
RQs

Can changes to trust in AS be detected & measured via behavioural cues & responses, physio & self-assessment?

Experts vs novices.

Technical Approach

BMT; trust in system modulated by auto classifier accuracy (25%, 75%, 95%)



Example Findings

- Trust plummets after cyber-attack & remains low in 25% & 75% conditions but not in 95% condition (restoration…)
- Evidence that subjective ratings do always correlate with objective physiological data!

Questions?

Centre for Artificial Intelligence, Robotics and Human-Machine Systems

Canolfan Deallusrwydd Artiffisial, Roboteg a Systemau Peiriant-Dynol

Prof Phil Morgan – *morganphil@cardiff.ac.uk*

School of Psychology

UNDEB EWROPEAIDD EUROPEAN UNION

Llywodraeth Cymru Welsh Government

Cronfa Datblygu Rhanbarthol Ewrop European Regional Development Fund

28th February 2024 – Bucharest, Romania

2023

IROHMS SIMULATION LAB

Cyber Security

Data Visualisation & C2

EEG, EMG, Eye Tracking, HSM

Igloo Immersive Dome

Robots (Nao, Pepper, TIAGO)

Transport Simulator

VR (Incl. Virtualizers)