Responsible research and development in AI for healthcare

Dr Kate Devlin

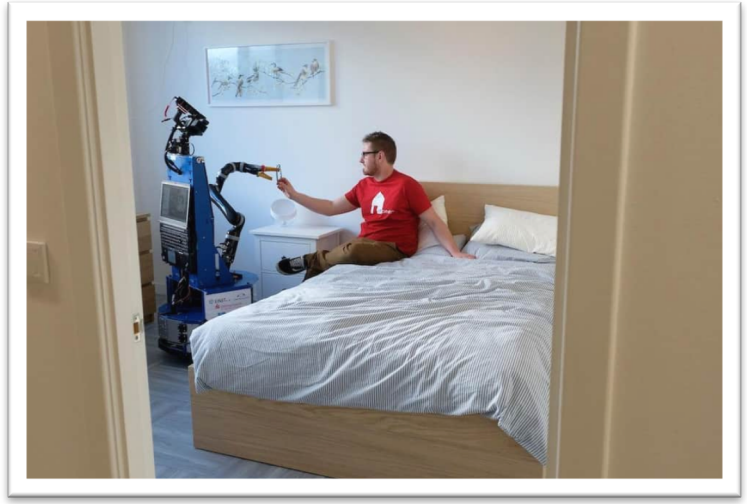
King's College London, UK.

UKRI Trustworthy Autonomous Systems Hub

KING'S College LONDON

Why do we need Responsible Research and Innovation (RRI) in AI for healthcare?

- Machine Learning is embedded and used in many of the applications we take for granted.
- AI shows clear promise in **automating diagnoses** and **personalising treatment**.
- However, global power inequalities in AI result in products that inherently benefit developers and often actively harm outgroups.
- Automation is not without its **social impact downstream**: autonomous systems can threaten clinicians' agency, could lead to reduced funding in staffing, and could result in job loss, as well as facing scepticism and concern from patients.
- Regulation is difficult and **there is no quick fix** to bias and a lack of transparency.



The Trusted Autonomous Systems Hub

- Focal point of the **£33mn UKRI TAS Programme**: UK government-funded research.

- Four-year project to “establish a collaborative platform for the UK to deliver world-leading best practices for the design, regulation and operation of 'socially beneficial' autonomous systems which are both **trustworthy in principle, and trusted in practice** by individuals, society and government.”

- **Three core universities** (Southampton, Nottingham, King’s College London) and **six more** as nodes that address functionality, resilience, security, governance and regulation, verifiability, and trust.

- Over **100 partners** (government, industry and NGOs). Partners must engage under a partnership agreement.

Remit of the TAS Hub

- We **fund projects**, set up **networks**, advise on **policy**, and invite researchers, industry, NGOs and the public to **engage** and **contribute use-cases/datasets** or **collaborate** on research projects, tech transfer, and training activities.

- TAS carries out research internally and also awards open grants to UK academics and industry partners.

- Subject areas include: autonomous vehicles (AVs), Net Zero, healthcare, maritime applications, art and creativity.

- Core principles of **responsible research and innovation** (RRI) with equal attention to **equality, diversity and inclusion** (EDI) – ethical by design.

- Fairness is centred in all that we do.

TAS Hub Priorities & Activities

UKRI Trustworthy Autonomous Systems Hub

Research	Advocacy & Engagement	Skills
Agile, Interdisciplinary, Industry Collaboration	Policy Makers, User Partners, Public	ECRs, Doctoral Training Network, Syllabus Lab

Creative Engagement & Media

THE NATIONAL GALLERY

Living With AI Podcast tas.ac.uk

BLAST THEORY

Industry Partners & Govt

Training & Outreach

What makes TAS research different?

Agile and Responsive

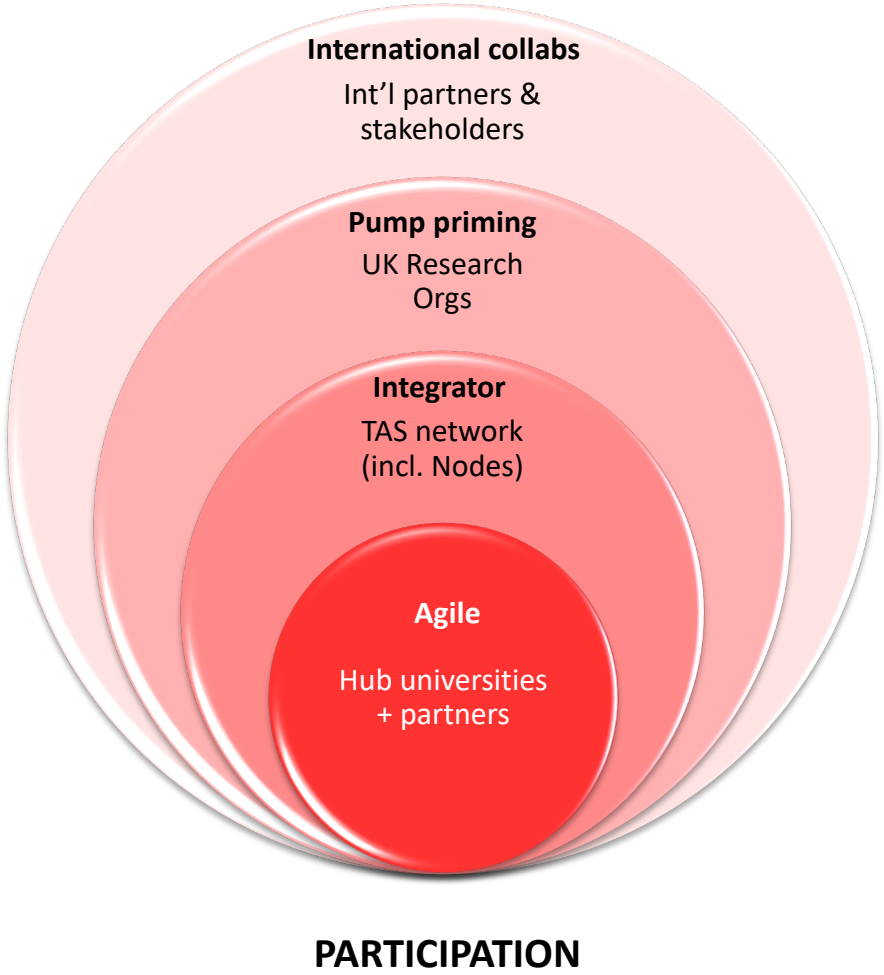
- Not predefined / prescribed / planned years in advance
- Annual cycles
- Current societal / technological concerns and opportunities

Participatory

- TAS community-led research ('bottom up')
- Stakeholder involvement (e.g., industry, users, policy makers)
- Multi-perspectival (disciplines, industries, views etc.)

Values-based (see tas.ac.uk/our-guiding-principles)

- Clue in the title: *trustworthiness* is a value
- Equality, diversity, & inclusion
- Responsible research and innovation



TAS Research Themes

- Initial analysis of the topics of the 12 pump priming projects and 6 Agile projects (confirmed by PP project leads)
- These projects cover a range of **disciplines**, examine different **issues related to trust** (e.g., facilitators vs. barriers)
- Across a range of **themes** and **research areas**

TAS RESEARCH AREAS

TAS TECHNIQUES AND MECHANISMS	PUBLIC ENGAGEMENT WITH AS	HUMAN-MACHINE INTERACTION	RESPONSIBLE INNOVATION PROCESSES	TAS GOVERNANCE

EXAMPLE TOPICS

➢ Transfer of control ➢ Flexible Autonomy ➢ Failure recovery ➢ Validation & Verification ➢ Resource allocation	➢ Perception of risks & opportunities ➢ Ethical concerns ➢ Explainable AS ➢ Human autonomy ➢ Acceptance, Adoption and participation	➢ Human-machine organisations ➢ Multi-agent Coordination ➢ Teaming & collaboration ➢ Managing conflicts ➢ Task sharing ➢ Social & societal impact	➢ Understanding stakeholder/ User needs ➢ Consent and Privacy ➢ Inclusion and Participation / PD ➢ Open science and Skills ➢ Human-centred design ➢ Production / financing	➢ Policy evaluation ➢ Resilience ➢ Regulation ➢ Legal implications ➢ Auditing

CROSS-CUTTING VALUES

Human-centredness; Equality, diversity & inclusion; responsible research and innovation; trust & trustworthiness; fairness, accountability, transparency & ethics (FATE)

How do we do this? Trustworthy IN PRINCIPLE:

- devising **criteria for grant reviews** that reward projects which centre stakeholder engagement;
- promoting **early career leadership** opportunities;
- ensuring **tangible ethical approaches**;
- writing **actionable equality, diversity and inclusion strategies** – and using them (e.g., a code of conduct);
- forming an **operational framework**;

- **collaborating with industry** (and researchers' varying reactions to who we work with – some people have strong views on who invests in our work); and
- the **choices we made when setting up** our Board, our Strategy Advisory Network and our International Scientific Committee.

How do we do this? Trusted IN PRACTICE:

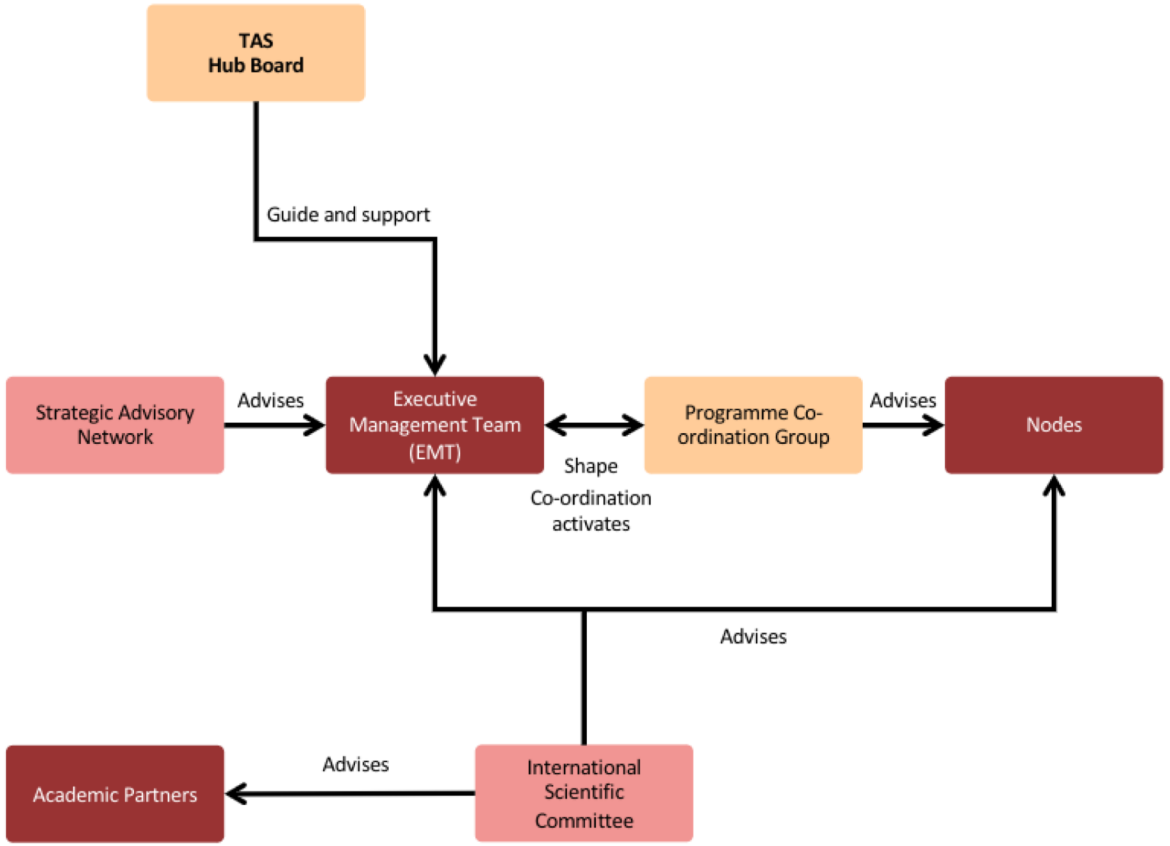
- Working with **stakeholders**, such as charities and mental health service-users, to explore ideas – e.g., around trust in digital mental healthcare systems;

- **Providing resources** such as video conversations, podcasts, and teaching materials; and

- Commissioning and developing **interactive creative artwork** that leads to thought-provoking encounters – for example, our Cat Royale art project that involves pets being cared for and played with by a robot arm .

What works well?

- Clear governance structure with allocated roles.
- Set specific and explicit aims early and revisit often.
- Oversight from independent boards/committees keeps us on track.
- All funding we award is assessed **FIRST** through the lens of RRI/EDI and supporting Early Career Staff, **THEN** rated on technical merit.



What has been difficult?

- **Occasional disagreement** over who we should partner with (e.g., some members unhappy with funding from areas such as Defence, or Big Tech).
- **Timescales**: some of our one-year pump-priming projects were ambitious and this meant some didn't meet the end of project deadlines.
- **Lack of integration**: we've found it harder to break into established non-academic circles, e.g., government; industry events. Perseverance and profile-boosting activities have helped.
- The **sheer scale of the project** means that we don't always have a full and comprehensive overview of all activities happening at any given time.
- We are **academically competing** with projects that centre results over responsibility.
- We **struggle to get creative**/non-STEM projects.

What we seek and what we can share

- We are keen to **collaborate and/or share ideas** with similar projects elsewhere. We have formed links with similar US projects working on socially beneficial AI:
  - UT Austin's Good Systems; Stanford's HAI group; Johns Hopkins' Institute for Assured Autonomy.

  We want to extend this outside of US/UK.

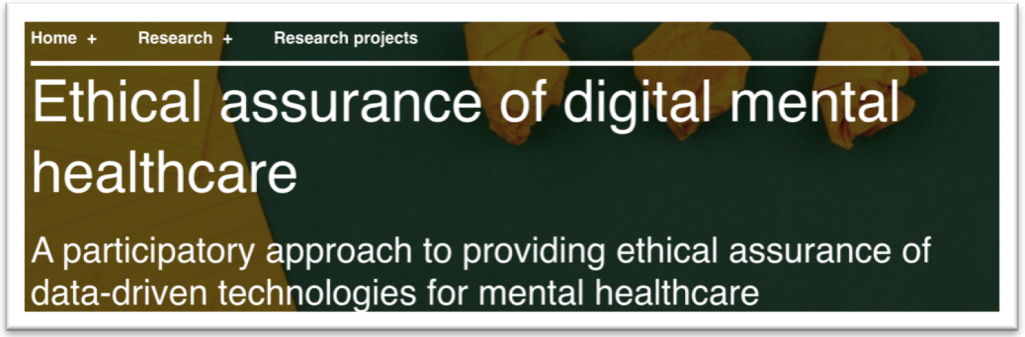
- We **share resources** on our website (https://www.tas.ac.uk/) including conference presentations, our podcast, and educational information.

- We are organising a **research symposium** to take place next July, open to anyone worldwide who works on the responsible development of AI and autonomous systems (in any discipline). We want to showcase and publish research that has put people and fairness at the core of the work.

Thank you

kate.devlin@kcl.ac.uk
Twitter: @drkatedevlin

contact@tas.ac.uk
Twitter: @tas_hub



Could a robotic device be trusted to heal your wounds using light?

Why will codesign methods be critical for acceptance and adoption of AI in chronic disease management?

Diagnostic AI System for Robot-Assisted A&E Triage

This collaboration between the Universities of York and Southampton and the York and Scarborough Teaching Hospitals aims to prototype a robot-assisted A&E triage solution for reducing patient waiting time and doctor workload.