UNIVERSITY OF BATH | INSTITUTE FOR POLICY RESEARCH

Policy brief

November 2026

Who is accountable when a digital twin gets it wrong?

Built-environment digital twins: Assigning responsibility for data, models and decisions heading

Dr Alejandro Jiménez Rios

Digital twins (DTws) – digital representations of real-world assets, interconnected through information about their condition and behaviour – help us to understand how physical assets work today, as well as to make better decisions about their future.[1] They are increasingly used in the built environment, providing advice on, and helping decision-making about, the management and operation of buildings, infrastructure assets and heritage sites.

Yet it is unclear who is accountable when an organisation permits autonomous action by a DTw, particularly one that results in or contributes to harm. Experts at the SafeDT Idea Factory event agree that DTws are only as trustworthy as the data and computational models used to create them. Their use brings risks, emerging from poor data, models that are unchecked against the real asset and overreliance on automated outputs.

Human users can affect levels of trust and limit reliability as well, particularly where there is a lack of skills or knowledge on how to use DTws. Users need to be aware of their limitations too. Participants from the SafeDT event treat these concerns as systemic failures, not technical glitches.

Government and standards bodies should therefore set minimum data requirements and clarify legal and operational accountability, with the strictest rules for the highest-risk uses. Asset owners should keep a named, competent person in charge of every safety-critical decision and check models against real assets regularly. Researchers, funders and professional bodies should develop practical safeguards and the necessary interdisciplinary skills to use them. Technology, people and governance must work together as no single measure or stakeholder are enough on their own.

1 Digital Twin: 'digital representations of target entities with data connections that enable convergence between the physical and digital states at an appropriate rate of synchronisation'. British Standards Institution (2026). *BS EN 18162:2026 Building information modelling (BIM) – Digital twins applied to the built environment – Concept and definitions*. London: BSI.

CoSeC Computational Science Centre for Research Communities

SafeDT

Introduction

In 2010, Stuxnet malware was used to target Iran's Natanz uranium enrichment plant. It drove centrifuge speeds up and down, destroying more than 1,200 machines. Meanwhile, normal readings were replayed to operators, meaning that technicians only spotted the fault by the sound of the malfunctioning equipment.[2] Similarly, in July 2026, hackers shut down a small UK power generator for four days.[3] Officials say the wider grid was unaffected, but the attack shows how quickly a digital weakness can become a physical infrastructure problem. Digital twins (DTws) sit at this junction between the digital and real worlds.

Many DTws primarily collect and visualise data at a descriptive level (Level 0). More advanced systems aggregate and analyse information, predict asset behaviour, or recommend interventions (Levels 1 to 3). At the highest maturity level (Level 4), a DTw may be authorised to initiate defined actions autonomously, subject to predetermined operational and safety constraints.[4] The UK government's National Digital Twin Programme (NDTP) aims to build national capability and make DTws safe, secure and trustworthy.[5]

DTws are used on public assets, such as bridges, roads, power plants and heritage sites, and on private ones, such as commercial buildings and utilities, to understand the assets' current state and to make predictions about their future performance and behaviour. The two often overlap: a private company may build, host or run the DTw of a public asset, and a private owner may operate an asset that the public relies on.

DTws are already in use in the UK. On the West Coast Main Line, fibre-optic sensors on new railway bridges near Stafford support a real-time DTw of a bridge, helping Network Rail and its partners to understand how it carries load and to monitor it as it ages.[6] On the Isle of Wight, a DTw built through the NDTP enables the council and emergency services to test how coastal erosion, landslips and flooding could cut off critical roads.[7] At Blenheim Palace, a heritage pilot combines centuries of archival records with live sensor data to predict risks to the historic estate and reduce energy use.[8] Some of these DTws serve public assets and some private ones, and most rely on partnerships between the two.

Challenges

Three pressures make the NDTP's aim for safe, secure and trustworthy DTws hard to meet.

First, artificial intelligence (AI) is becoming more capable and autonomous. For example, US AI corporation Anthropic reports that its Claude Mythos preview model finds and exploits software weaknesses better than all but the most skilled humans, which helps defenders and attackers alike.[9] This potential is already evident in practice: in October 2026, cybersecurity researchers used AI-assisted analysis to identify two critical vulnerabilities in PHP, a programming language underpinning a large proportion of the internet. The PHP2Shell vulnerabilities had remained undetected for years and could enable attackers to take control of affected web servers, prompting an international cybersecurity response. This also demonstrated how AI can help to uncover weaknesses that might otherwise escape human detection.[10] Because a DTw links sensors, software and control systems, it gives such tools many points to probe, and a successful attack could feed it false data or change what it tells operators (as in the Stuxnet attack).

2 Control Engineering (2021). Throwback attack: how Stuxnet changed cybersecurity
3 The Telegraph (2026). Iranian hackers shut down UK power plant, 22 August.
4 Palmer, C. et al. (2025). Personas to inform cognitive interaction with a digital twin. *Journal of Industrial Information Integration*, 47, p.100893.
5 Department for Business and Trade (2023). National Digital Twin Programme (NDTP)
6 Global BIM (2025). Staffordshire bridges – long term performance monitoring using fibre-optic sensors
7 National Digital Twin Programme (2023). Asset resilience, emergency planning and response: the National Digital Twin Programme (NDTP). GOV.UK.
8 Cesium (2026). Blenheim Palace Performing Smart Conservation with Cesium.
9 Anthropic (2026). Project Glasswing: securing critical software for the AI era, 7 April.
10 BBC News (2026). What is PHP and why did it spark a global cyber-security scramble?, 2 November.

Second, experts put the chance of a quantum computer being able to break today's encryption within 10 years at 28-49%.[11] A DTw holds records for an asset's whole life, so data protected by today's encryption could become vulnerable once such technology is deployed in practice. This risk is compounded by so-called 'Harvest Now, Decrypt Later' (HNDL) attacks, in which malicious actors collect encrypted data today with the intention of decrypting it once sufficiently powerful quantum computers become available. Further, the threat is not limited to the confidentiality of historical records: compromised cryptographic systems could also undermine the integrity and availability of DTw data, potentially affecting the reliability of monitoring, decision-making and operational control.

Third, the 2026 encyclical *Magnifica Humanitas* argues that responsibility and meaningful human judgement must remain with people as AI spreads.[12] As DTws move from advising (Levels 0 to 3) to deciding (Level 4), this makes it essential to define where a person must stay in charge of decisions about safety-critical assets.

Methodology

DTws now inform decisions about the management and operation of critical infrastructure, buildings and heritage sites. This policy brief asks who is accountable when those decisions go wrong. It draws on the outcomes from an Idea Factory Event organised within the context of a Computational Science Centre for Research Communities (CoSeC)[13] Fellowship research project titled Safe Digital Twins for the Built Environment (SafeDT).[14] Experts from academia, industry and the government were invited to participate to map the risks of DTws and propose adequate safeguards. The policy brief sets out what they found and what government, researchers, industry and asset owners should do next.

The consultation on 18 September 2026 involved 33 participants, predominantly from UK universities, with additional industry and government perspectives. Participants used a live poll and were divided into four groups to work on digital boards during the collaborative sessions of the event. Findings should therefore be treated as an expert scoping exercise rather than a representative survey of asset owners, operators, regulators, technology suppliers or affected communities. That said, participants agreed on one point above all: a DTw is only as trustworthy as the data, the models and the people behind it.

Findings

Poor data weaken everything that follows

The risk
Errors travel. Faulty sensors, missing readings, out-of-date records and inconsistent definitions produce unreliable models. These in turn create misleading results and lead to poor decisions. As one participant stated: 'rubbish in – rubbish out'.

Participants ranked data quality as the most serious concern. Further, infrastructure lasts far longer than software, so one group asked whether today's data will still be readable in 50 or 100 years. This means that a DTw needs data and models that are kept up to date throughout the asset's life, not only checked at launch.

Computational models and processes will change completely over the life of an infrastructure asset or significant building. It is hard to believe that a piece of water infrastructure designed and built to last 60 years (and with an expected life of longer than that) will not see numerous different approaches to gathering and using its data over that time. This also means planning for longevity from the start, by recording what data are telling us as well as how they were collected (in other words, high-quality metadata are key).

11 Mosca, M. and Piani, M. (2026). *Quantum threat timeline report 2025*. Global Risk Institute.
12 Leo XIV (2026). *Magnifica Humanitas: on safeguarding the human person in the time of artificial intelligence*. Vatican City: Vatican.
13 CoSeC: Computational Science Centre for Research Communities
14 Safe Digital Twins for the Built Environment

What participants propose
Common standards for how data are collected, checked, updated and shared. Regular validation of data. Open practices, so that people know what model is being used for what purpose and can see what data sit behind a model. Methods to measure the quality and accuracy of data.

Who acts
Government and standards bodies set the standards and guidance on data sharing, while asset owners and operators check and update data throughout an asset's life. Researchers develop ways to measure data quality and uncertainty.

A sophisticated model is not a necessarily a trustworthy one

The risk
A model can look impressive yet still be wrong. Participants worried about models that are not checked against the real asset, 'black box' systems whose reasoning nobody can explain, and weak understanding of how the physical structure behaves. A DTw with outputs that cannot be trusted gives no sound basis for decisions.

What participants propose
Regular checks against real measurements, and recalibration as the asset and its conditions change. For higher-risk uses, independent verification and clear reporting of uncertainty.

Who acts
Asset owners and operators validate models against real data. Technology developers build in a deep understanding of the physical asset and keep the model in step with it. Researchers develop better ways to represent how the physical asset behaves, quantify and communicate uncertainty, and explain how conclusions are reached.

People must stay in charge

The risk
One participant highlighted a risk shared by many: 'My biggest concern is people trusting a DTw too much'. As DTws move from showing data to advising and deciding, people may stop questioning them, or lose the practical understanding needed to spot wrong answers.

What participants propose
- **Human intervention before a decision.** It should be settled in advance when a person must step in. Where a DTw advises, a competent person makes the final decision, with checkpoints at which someone must approve before action. Where a DTw can act on its own, its limits must be set beforehand: which actions it may take, which need approval, and who can stop it. Staff need training to help them question the system and know when to escalate.
- **Accountability after a decision.** Where a person decides to use a DTw, that person and their organisation remain accountable. In addition, the DTw's advice should be recorded so that decisions can be reviewed. Where a DTw acts autonomously, a named owner must remain accountable for allowing it to act. Participants also noted that responsibility may be distributed across multiple actors, including data providers, model developers, platform providers, asset owners and operators. Thus, participants ask government to define this accountability, because the answer may differ in each case.

Who acts
Asset owners and operators put human sign-off in place. Government sets the rules: who is accountable, who has authority to intervene and how far a DTw can be relied on. Researchers study how people and automated systems work together.

Accountability is still unclear

Participants repeatedly asked who is accountable when a DTw contributes to a harmful decision. They want clear legal responsibilities, but flexibility depending on different cases. For example, a DTw for a low-risk building should not face the requirements of those for critical infrastructure.

The groups did not agree with a specific legal answer, and facilitators saw only differences of emphasis, such as how to balance data sharing with data protection. Not every aspect of the DTw operation can be governed in every context by law; some will need to be asset/DTw-specific, and this can only realistically be done through agreements between parties. There is also a relationship between stakeholders and organisational policy/capability to implement these operationally.

What the law says now
The UK has no law written for DTws and no AI-specific statute. The government relies on existing regulators and general principles.[15] Responsibility therefore falls under rules written for other purposes.

The Building Safety Act 2022 requires a 'golden thread' of accurate information and named accountable people for higher-risk buildings.[16] The Data (Use and Access) Act 2025 gives people information about, and the right to challenge and seek human review of, significant decisions made solely by automated means about them. It covers decisions about people, not assets.[17]

At the time of writing, the Cyber Security and Resilience (Network and Information Systems) Bill is progressing through the House of Lords. It is intended to strengthen and extend requirements governing the resilience of network and information systems used for essential activities and relevant digital services.[18]

At present, when several parties supply data to a DTw, it is hard to show who caused an error, and none of these rules set out who answers when a DTw contributes to harm. Further, complex supply chains, shared data environments and evolving models can make causation, control and responsibility difficult to evidence.

What should change
Government should first review the extent to which existing health and safety, building safety, professional liability and procurement rules cover decisions informed or taken by DTws, and publish the gaps. Where gaps exist, it should set requirements that scale with risk – for example, for safety-critical assets, a named accountable person, a record of what the DTw advised and clear authority to intervene.

Public bodies should also use contracts to settle who owns the data and model, and who is responsible for errors, when a private firm builds or runs the DTw of a public asset. Technology developers should also build traceability into DTws. This means keeping a record of where each data input came from, which version of the model produced each output and what the DTw advised, so that the cause of an error can be traced, particularly when several parties are involved.

Other concerns

Cost and value matter. Participants fear DTws that are expensive to maintain and cannot show a return. Indeed, one significant economic question that is yet to be fully explored is who benefits from any returns. Several warned that DTws are oversold, and even that the term 'is becoming toxic because of over and improper use'.

Concerns also remain over whether DTws are built around what is technically possible rather than what users need. They ask for DTws to be co-designed with the people who will use them.

Security also worried participants, from attacks using AI to false or manipulated sensor data and failures in critical national infrastructure. Yet no group ranked this among its top priorities, which suggests it needs closer attention.

15 RPC (2026). Part 1 – UK AI regulation, AI guide, 1 June.
16 Building Safety Act 2022, c. 30.
17 Department for Science, Innovation and Technology (2025). Data (Use and Access) Act 2025: data protection and privacy changes. GOV.UK.
18 UK Parliament (2026). Cyber Security and Resilience Bill

Conclusions

DTws are advancing. They are going beyond showing data to advising on, and sometimes making decisions autonomously, about the management and operation of buildings, infrastructure assets and heritage sites.

As this shift occurs, the rules governing DTws need to be revised to keep up, as highlighted by the experts consulted for this policy brief. For example, one group warned that technology develops far faster than the governance needed to apply ethics and regulation. Another argued that DTws risks should not be treated as technical glitches, but as 'systemic failures requiring proactive governance and structural guardrails'.

This view matters because the pressures described in the introduction are growing. More capable and autonomous AI, cyberattacks that reach physical assets, and the longer-term quantum computing threat to today's encryption all raise the cost of a wrong or manipulated DTw output. A manipulated DTw could hide damage from operators, as in the case of the Stuxnet cyberattack.

Accuracy and sound use are vital, including reliable data, a model checked against the real asset and a competent person in charge. These determine whether a DTw gives the right answer and whether people act on it wisely. Indeed, participants saw these as the foundation for everything else.

Security is a different concern, one focused on protecting a DTw from attack or manipulation. Participants raised this, but they rank it lower than other risks. Security protects accuracy as manipulated data produce wrong outputs. This suggests that closer attention needs to be paid to security before DTws take on more responsibility and are allowed to act fully autonomously.

The central policy challenge is accountability. The UK government aims to make DTws safe, secure and trustworthy. Participants highlighted what this requires in practice: clear standards for data, regular validation of models against real assets, human sign-off on safety-critical decisions and clearly defined responsibility when something goes wrong. They also called for proportionate rules, with the strictest requirements reserved for the highest-risk uses, such as assets where a wrong or manipulated output could cost lives, cut essential services or cause irreversible loss. Examples include hospitals, higher-risk buildings, bridges and critical infrastructure and heritage sites.

No single measure is enough. Data standards without skilled people to question results, or human oversight without a validated model, leave gaps that cause failures. Technology, people and governance have to align, and the actors who hold each element – from government to asset owners, researchers and industry – need to act together.

These conclusions rest on a consultation made up almost entirely of academics. Testing them with industry, government and asset owners is the next step, and the recommendations that follow set out where to start.

Recommendations

Government and standards bodies should set minimum data requirements for DTws that support safety-critical assets.

Standards should cover how data are collected, checked, updated and shared, and should protect against false or manipulated inputs. They should build on the new UK and European standard, BS EN 18162:2026, that defines DTws in the built environment.[19] This gives industry a shared language for what a DTw is, how its quality can be specified, and is intended as a base for further standards.

19 British Standards Institution (2026). *BS EN 18162:2026 Building information modelling (BIM) – Digital twins applied to the built environment – Concept and definitions*. London: BSI.

It does not itself set requirements for checking data, security or accountability, and like most UK standards it is voluntary unless a contract or regulation requires it. Consequently, the government should also publish simple guidance on what data to share, with whom and how to protect privacy.

First step: Start with guidance for public and infrastructure assets, then move towards required standards.

Government should clarify who is legally accountable when a DTw contributes to a harmful decision.

There is currently no clear process or hierarchy for accountability. Rules should set out who is responsible and who has authority to step in, and they should scale with risk: stricter for a bridge or hospital than for a low-risk building.

First step: Review whether current health and safety, professional liability and procurement rules already cover decisions informed by DTws and publish the gaps.

Asset owners and operators should keep a named, competent person in charge of every safety-critical decision.

Asset owners and operators should establish explicit organisational accountability and named, competent roles for every safety-critical DTw application. They should document whether the system informs, recommends, controls or acts; define who may approve, override or suspend its operation; and preserve an auditable record of the data, model version, uncertainty and advice supporting significant decisions.

First step: Add human sign-off and a clear statement of the DTw's role to existing asset management procedures.

Researchers and research funders should prioritise practical safeguards.

Participants identified gaps in measuring data quality and uncertainty, explaining how AI-supported conclusions are reached, understanding how people use these systems in real work and which models are suitable for which function, and testing DTws against cyberattacks.

First step: Fund projects that work with asset owners to test these methods on real assets.

Professional bodies and universities should build interdisciplinary skills.

Engineers, operators, data specialists and developers need to understand each other's work, and users need confidence to question a DTw's output.

First step: Develop shared training and professional guidance and consider competence requirements for safety-critical uses.

Acknowledgements

This work was supported by the Computational Science Centre for Research Communities (CoSeC) 2025 CoSeC Fellowship that funded the SafeDT project (1 October 2025 to 31 December 2026). The SafeDT Project Lead thanks the Institute for Policy Research at the University of Bath for its support in preparing this policy brief, and the participants, keynote speakers and PhD student facilitators of the SafeDT Idea Factory event for their time and insights.

Project lead

Dr Alejandro Jiménez Rios
SafeDT Project Lead, Department of Architecture and Civil Engineering, University of Bath
Email: ajr225@bath.ac.uk

Contributing authors

Monika Negi, Department of Architecture and Civil Engineering, University of Bath
Yufei Dai, Department of Architecture and Civil Engineering, University of Bath
Salma Ahmed, Department of Architecture and Civil Engineering, University of Bath
Tianyu Xiang, Department of Architecture and Civil Engineering, University of Bath
Thomas Beach, School of Engineering, Cardiff University
Ali Rahman, Faculty of Engineering and Physical Sciences, University of Leeds
Gürkan Gür, Zurich University of Applied Sciences (ZHAW)
Leon Black, Faculty of Engineering and Physical Sciences, University of Leeds
Yingpei Sun, Department of Engineering, Durham University
Wei Zhang, Department of Engineering, University of Exeter
Petia Tzokova, Department of Engineering, University of Cambridge
Simon De-Ville, Department of Civil and Environmental Engineering, University of Liverpool
Zhiqi Hu, School of Architecture, Building and Civil Engineering, Loughborough University
Mohammed Abdelmegid, Faculty of Engineering and Physical Sciences, University of Leeds
Gloria Castro, Department of Civil Engineering, University of Birmingham,
Mahmoud Abdelrazek, Advanced Research Computing, University College London
Ricardo Codinhoto, Department of Architecture and Civil Engineering, University of Bath
Alhadi Alsaffar, School of Engineering, Cardiff University
Paul Emeka Eke, Faculty of Engineering and Physical Sciences, University of Leeds
Krishnan Guruvayur Venkateswaran, Department of Civil Engineering, University of Birmingham
Serik Tokbolat, Department of Civil Engineering, University of Nottingham
Mujib Adeagbo, School of Engineering, Cardiff University

Contact the Institute for Policy Research

Institute for Policy Research
University of Bath
Claverton Down
Bath
BA2 7AY

ipr@bath.ac.uk

Visit our website
bath.ac.uk/ipr

Read our blog
blogs.bath.ac.uk/iprblog

Follow us on LinkedIn
linkedin.com/school/bath-ac-uk-ipr

Follow us on Bluesky
@UniofBathIPR.bsky.social

Join our mailing list
bit.ly/IPR_newsletter