Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predictive governance of agentic city digital twins through a Decision Twin architecture grounded in Artificial Complementary Intelligence

Mahdi Seify

mahdi.seify@northampton.ac.uk

University of Northampton

Research Article

Keywords: Digital twins, Decision twin, Agentic AI, Urban governance, Artificial Complementary Intelligence, Predictive control, Decision rights, Smart cities

Posted Date: November 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11502869/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

City digital twins have become proficient at mirroring urban systems but remain weak at governing the decisions taken through them. As agentic artificial intelligence (AI) begins to act on urban infrastructure, two gaps widen: a control gap, because twin outputs are seldom linked to the strategic targets of the institutions that own infrastructure, and a legitimacy gap, because no principled rule allocates decision authority between machines, accountable officials and citizens. This article proposes the Decision Twin, a digital twin extended with four governance layers: predictive feedforward control based on Deep Business Analytics (DBA); a nested Business Connectivity Map (BCMap) that composes node-level forecasts into city-level indicators; an urban decision-rights function grounded in Artificial Complementary Intelligence (ACI); and a Ulysses Protocol of precommitments that the agentic layer cannot override. The design is developed through design science research and demonstrated by secondary analysis of a completed doctoral case: a sensor-instrumented haematology laboratory in an NHS hospital, treated as a node of critical urban infrastructure. More than 10 million two-minute records over 28 days were linked to 52 performance indicators. A human-weighted, disaggregated configuration forecast most accurately (RMSE 0.00436, against 0.0090 for a single pooled model). A retrospective audit found that only 3 of 12 governance decisions were computational. A day-ahead forecast showed an energy saving that signalled a service shortfall, illustrating how siloed sector twins misread cross-sector effects. The article also contributes an Extrapolation Escalation Rule, a Decision-Rights Drift Index and a maturity pathway for city twins.

1 Introduction

Cities are complex adaptive systems in which energy, mobility, health, water and housing interact in ways that no single agency observes in full. Digital twins promise a remedy: a virtual representation of physical urban systems, updated from sensors in real or near-real time, that offers a shared view across sub-systems and supports prediction [1, 2]. Research on city-scale twins has grown rapidly, moving from three-dimensional city models towards integrated platforms for planning, operations and participation [3–6], with operational examples ranging from Zurich's planning twin to Herrenberg's participatory model [7, 8]. Recent systematic reviews and an expert survey nevertheless report a fragmented concept, persistent interoperability and data problems, uncertain practical value, and governance, organisational and social issues that are often overlooked [9–11].

Most of what is called a city digital twin, however, still mirrors rather than steers. Kritzinger et al. distinguish a digital model, in which data flow manually in both directions, a digital shadow, in which data flow automatically from the physical to the virtual object only, and a full digital twin, in which data flow automatically in both directions [12]. By that standard, many urban deployments are shadows: they ingest sensor data and present it on dashboards, while decisions are taken elsewhere, through procedures the twin does not see [13, 14]. Socio-technical studies of city-scale twins have shown that the hardest problems are not computational but organisational: purpose, ownership and accountability [15].

Agentic AI changes the stakes. Agentic systems pursue goals, plan and act with limited human supervision, and their potential for harm increases as they become more agentic [16]. Urban AI is already moving from automation towards autonomy, with AI systems taking decisions in and about urban space [17, 18]. Harari observes that AI is the first technology able to make decisions and generate new ideas by itself [19]. Once a twin can retime traffic signals, shift energy loads or reschedule maintenance on its own, the question is no longer how faithfully it mirrors the city but who decides through it.

Two gaps follow. The first is a control gap. Management control systems, including the Balanced Scorecard, were designed around retrospective feedback: performance is measured, compared with targets and corrected after deviations have occurred [20–22]. City twins generate predictions, yet these predictions are seldom wired into the strategic performance architectures of the institutions that own urban infrastructure, so a forecast rarely becomes a pre-emptive decision. The second is a legitimacy gap. Ethical principles for AI are widely published but do not by themselves determine who may decide what [23]. Human operators over-trust automated recommendations [24, 25], public decision-makers tend to adopt advice selectively when it matches existing stereotypes [26], and requirements for human oversight often fail in practice, sometimes lending legitimacy to the systems they are meant to check [27]. The European Union's AI Act classes AI used as a safety component in the management and operation of critical digital infrastructure, road traffic and the supply of water, gas, heating and electricity as high-risk, and requires such systems to be designed for effective human oversight [28]. The Act requires oversight; it does not specify which urban decisions a machine may take, which require an accountable official, and which belong to citizens.

This article addresses both gaps by proposing the Decision Twin: a digital twin extended with governance layers that convert prediction into pre-emptive, accountable decisions. The design builds on the author's doctoral development of Deep Business Analytics (DBA), an AI-enabled management control framework that adds predictive feedforward control to the Sustainable Balanced Scorecard [29], and on Artificial Complementary Intelligence (ACI), a boundary framework that specifies where algorithmic authority ends [30]. Three research questions guide the work:

RQ1 (control)

How can a city digital twin move from state replication to pre-emptive, indicator-linked control, and does the doctoral evidence show this is feasible at node level?

RQ2 (legitimacy)

What principled rule should allocate decision rights between agentic AI, accountable officials and citizens in city-scale twins?

RQ3 (scale and accountability)

How can node-level control and decision rights be composed to city scale while keeping human judgement measurable?

To the author's knowledge, no existing framework links city twin forecasts to institutional performance targets while also allocating decision authority across machines, officials and citizens. The article makes five contributions. First, it defines the Decision Twin and reframes the return path from virtual to physical systems as a governance question rather than a question of technical maturity. Second, it formalises a nested BCMap that composes node-level forecasts into institutional, sector and city indicators, with explicit cross-sector coupling. Third, it extends the ACI decision-rights function to urban settings by adding reversibility, equity exposure and envelope status, and derives an Extrapolation Escalation Rule. Fourth, it proposes a Ulysses Protocol of democratic precommitments and a Decision-Rights Drift Index (DRDI) that makes the erosion of human authority measurable. Fifth, it demonstrates the design on a sensor-instrumented NHS laboratory, using only data collected during the doctoral research. These contributions speak to the Collection's themes of ontology-driven and multi-resolution twin architectures, agentic AI and participatory decision-making, ethics and governance, and standards for integrated urban ecosystems.

Section 2 reviews the theoretical background. Section 3 describes the research design. Section 4 presents the Decision Twin architecture. Section 5 reports the evidence from the instrumented node. Section 6 discusses contributions and implications, Section 7 sets out limitations, and Section 8 concludes.

2 Theoretical background

2.1 City digital twins: from state replication to agency

Grieves and Vickers describe a digital twin as three elements: a physical space, a virtual space and the connection that carries data and information between them [1]. The ISO 23247 framework, written for manufacturing but widely borrowed, organises a twin into an observable physical domain, a device communication domain that collects data and controls devices, a digital twin domain that hosts models and services, and a user domain [31]. The ISO/IEC 30173 vocabulary now standardises these concepts across domains [32]. Batty notes that cities stretch these definitions, because a city is not an engineered product with a design specification but a system of systems shaped by human behaviour [2]. He has since argued that the gap between a twin and its referent is wider for social and organisational systems than for physical assets, that cities may be inherently unpredictable, and that humans must therefore remain in the loop in the design and use of urban twins [33].

Reviews of city-scale twins describe a concept still without an agreed definition, extending from three-dimensional city models towards semantic data, real-time sensing, physical models and simulation [3, 4, 9]. Deng et al. position the twin city as a new pattern of urban governance [5], and case studies such as the West Cambridge campus twin show that building- and city-level twins can integrate heterogeneous data to support decisions in operation and maintenance, the longest phase of the asset life cycle [34]. The Gemini Principles for the United Kingdom's national twin programme insist that twins must have a clear purpose, be trustworthy and function effectively [35], and Nochta et al. argue that city twins must

be understood as socio-technical systems whose value depends on how they are embedded in institutions [15]. Charitonidou observes that urban-scale twins promise wider citizen participation in planning decisions but rest on a limited set of variables and processes, which makes that promise problematic [36], while citizen-feedback twins show that public input can be built into the platform itself [37]. Security research adds that centralised twins create new attack surfaces in the physical systems they mirror [38].

Across this literature the return path from virtual to physical is treated as a matter of automation maturity: a twin is more advanced the more of the loop it closes automatically [12]. The CITYSTEPS maturity model, for example, places real-time and then autonomous decision-making and implementation at the top of its eight stages [39]. Korenhof et al. offer a different reading: a twin is less a neutral copy than a technique for steering its physical counterpart, which makes the questions of who sets its goals and who holds the wheel central [40]. The automation-maturity framing becomes unsafe once the twin hosts agentic AI. The more relevant question is which return flows should be automated, which should pass through accountable officials, and which should pass through citizens. That is a question about decision rights, and the twin literature does not yet answer it.

2.2 Management control and predictive feedforward

Management control systems are the mechanisms through which organisations translate strategy into action and correct deviations [22]. Simons identified four levers of control: belief systems, boundary systems, diagnostic control systems and interactive control systems [20]. Malmi and Brown argued that these controls operate as a package, combining cybernetic, administrative and cultural elements [41], and Ferreira and Otley extended the analysis to the design and use of whole performance management systems [42]. The Balanced Scorecard operationalises diagnostic control across financial, customer, internal-process and learning perspectives [21], and its sustainability extension adds environmental and social objectives [43]. In cybernetic terms, a controller must have at least as much variety as the system it regulates [44]; a scorecard reviewed monthly has far less variety than a city that changes minute by minute.

All four levers share a temporal limitation: they process information about events that have already occurred [30]. Urban performance systems inherit this limitation. City indicators, benchmarks and dashboards are largely diagnostic, reporting what has happened rather than what is about to happen [14]. Standardised indicator sets such as ISO 37122 define what smart cities should measure [45], but not how forecasts of those indicators should trigger decisions or who should take them. Deep Business Analytics was developed to address it. In the author's doctoral research, DBA linked sensor and operational data, through a Business Connectivity Map (BCMap) of cause-and-effect relationships, to a Sustainable Balanced Scorecard, and used long short-term memory (LSTM) networks [46] to forecast key performance indicators one step ahead [29]. The difference between a forecast and its target becomes a feedforward signal that allows managers to intervene before a deviation occurs. Seify formalised this as a fifth lever of control, Predictive Feedforward Control [30]. City twins already produce

forecasts; what they lack is the control architecture that turns forecasts into accountable, target-linked decisions.

2.3 Artificial Complementary Intelligence, complementary assets and automation bias

The term artificial intelligence invites the assumption that machines replicate human intelligence as a whole. Gardner's theory of multiple intelligences rejects a single general intelligence and identifies distinct capacities, including linguistic, logical-mathematical, spatial, interpersonal and intrapersonal intelligence [47]; he later discussed existential intelligence, concerned with questions of meaning and purpose, as a candidate capacity [48]. Current AI systems are strong in pattern recognition and logical-mathematical processing and weak, or absent, in the interpersonal, intrapersonal and existential capacities on which accountable judgement depends. ACI takes this asymmetry as a design principle: algorithms are complements to human judgement, not substitutes for it [30]. In health governance, ACI specified a four-domain taxonomy in which computational decisions may be delegated under oversight, contextual decisions are AI-assisted but human-determined, and ethical and relational decisions remain exclusively human [30].

Strategic management theory reaches a similar conclusion from a different direction. Resources confer sustained advantage only when they are valuable, rare, inimitable and non-substitutable [49]. Algorithms and platforms commoditise quickly, so the value of a technology is captured by the complementary assets that surround it [50]. Collective-intelligence and hybrid-intelligence research adds that the best performance often comes from people and computers organised to think together, not from either alone [51, 52]; in organisational decisions, AI handles complexity well while people remain better at uncertainty and equivocality [53]. Raisch and Krakowski show, however, that automation and augmentation are interdependent: augmentation can turn into automation over time, and an exclusive focus on either harms organisations and society [54].

Human factors research explains why complementarity does not maintain itself. Operators misuse automation through over-reliance [24]; automation bias persists even when recommendations are demonstrably wrong [25]; levels of automation can be set separately for information acquisition, analysis, decision selection and action implementation [55]; and the more a system is automated, the less practised its human supervisors become at the moments when they are most needed [56]. A further limit is statistical. Machine-learning models interpolate well within the distribution of their training data but extrapolate poorly beyond it; when inputs shift, both accuracy and the reliability of the model's own uncertainty estimates degrade, often without warning [57–59]. Rudin argues that in high-stakes settings interpretable models should be preferred to black boxes, and that the presumed trade-off between accuracy and interpretability is often absent [60].

2.4 Institutions, precommitment and civic legitimacy

Institutions are the rules of the game that structure human interaction [61], and governance structures exist to economise on bounded rationality while safeguarding against opportunism [62]. Algorithmic governance needs institutions in this sense: rules that persist when convenience argues for their relaxation. Elster's analysis of precommitment uses Homer's Ulysses, who had himself bound to the mast so that he could hear the Sirens without steering towards them [63]. Societies use the same device when they entrench constitutional rules against majorities of the moment [64]. Laux argues in the same spirit that human oversight under the AI Act should be designed around institutionalised distrust, the democratic tradition of checks that do not depend on the goodwill of those being checked [65]. Precommitment is relevant to city twins because the pressure to extend automation is continuous and local: each step is efficient, while the cumulative drift of authority from people to machines is rarely decided at all. Ethical principles alone do not resist this drift [23]; enforceable institutional constraints can.

Public administration research has begun to specify which tasks can be delegated to artificial discretion. Bullock treats task complexity and uncertainty as the defining criteria for whether a task favours human discretion or AI [66], and Young et al. assess artificial discretion against task specificity and environmental complexity, finding potential gains in scalability, cost and quality alongside serious concerns about equity, manageability and political feasibility [67]. Busuioc shows that opacity and automation bias weaken officials' ability to hold algorithms to account [68], and Danaher warns that decision-making by algorithm can erode the legitimacy of public decisions if citizens cannot understand or contest them [69]. Santoni de Sio and van den Hoven define meaningful human control by two conditions: the system must respond to the relevant reasons of the people concerned, and its outcomes must be traceable to a person who understands the system and accepts responsibility [70].

Urban decisions also differ from organisational ones because their ends are contested. Planning problems are wicked: they have no definitive formulation, and their solutions are judged good or bad rather than true or false [71]. Participation ranges from manipulation to citizen control [72]; in smart-city programmes, most initiatives described as citizen-centric rest on stewardship, civic paternalism and a consumer view of citizenship rather than on rights and the common good [73], and smart urban governance depends on socio-technical collaboration rather than technology alone [74]. Rahwan proposes society-in-the-loop governance, in which the values encoded in algorithmic systems are negotiated with the public rather than fixed by engineers [75]. Automated public systems can also concentrate their burdens on poorer and more vulnerable groups [76]. Maslow's later work placed self-transcendence, a commitment to purposes beyond the self, above self-actualisation [77, 78]. At city scale, the corresponding question is what the city is for. That question is the core of civic deliberation, and no optimisation model can answer it, because it sets the objective function that optimisation assumes.

2.5 Synthesis

Table 1 summarises what each body of work offers and leaves open. Taken together, they suggest that a city twin able to act needs three things it does not yet have: a control architecture that turns forecasts

into target-linked decisions; a principled allocation of decision rights that reflects both cognitive capacities and civic legitimacy; and institutional safeguards that hold that allocation in place as automation grows.

Table 1
Literature streams, open problems and the Decision Twin response

Literature stream	What it offers	What it leaves open	Decision Twin response
City digital twins [10, 15, 39]	Data integration, visualisation and simulation across urban sub-systems	The return path is treated as automation maturity, not as a governance choice	Governed return path through explicit decision rights
Management control systems [20, 21, 41]	Translation of strategy into targets, measures and corrective action	Retrospective feedback; no link to real-time urban forecasts	DBA predictive feedforward and nested BCMap
AI boundary frameworks [30, 47]	Principled limits on algorithmic authority in health governance	Static domains; no treatment of reversibility, equity or distribution shift	Urban decision-rights function and Extrapolation Escalation Rule
Human–automation interaction [24, 26, 27, 56]	Evidence of automation bias and skill erosion	Few continuously computable organisational measures	Decision-Rights Drift Index alongside the Judgment Preservation Index
Institutions, public administration and participation [63, 65, 66, 75]	Precommitment, artificial discretion, institutionalised distrust and society-in-the-loop governance	Not yet translated into twin architectures	Ulysses Protocol and a civic–deliberative domain

3 Research design

3.1 Design science approach

The study follows design science research, which produces and evaluates artefacts intended to solve identified organisational problems [79]. Peffers et al. describe six activities: problem identification, definition of objectives, design and development, demonstration, evaluation and communication [80]. In this study, the problem is the control and legitimacy gap of agentic city twins (Sections 1 and 2); the objectives are the three research questions; the artefact is the Decision Twin architecture (Section 4); the demonstration and evaluation use a completed doctoral case (Section 5); and this article is the communication activity. The artefact comprises constructs (Decision Twin, decision-rights domains, envelope status), models (nested BCMap), methods (Extrapolation Escalation Rule, Ulysses Protocol) and a measure (DRDI), the four artefact types recognised in design science [79]. In Gregor and Hevner's terms the contribution is primarily an exaptation, extending solutions proven in organisational control

(DBA) and health governance (ACI) to the new problem of agentic city twins, with improvement elements where new constructs are required [81]. A generative AI tool, Claude (Anthropic), was used only to support drafting and language editing, as described under Declarations; it was not used to design the study or to generate data.

3.2 Case and data

The demonstration uses a single embedded case [82]: the Haematology and Coagulation Laboratory at the Royal Liverpool Hospital (HL@RHL), where the author developed and validated DBA during doctoral research [29]. The laboratory is treated here as a node of critical urban infrastructure. Hospital diagnostic services sit at the intersection of the health and energy systems of a city, and the laboratory was instrumented much as a facility within a city twin would be. A single case is appropriate because the aim is to show that the architecture can be instantiated and to expose its mechanisms, not to estimate population effects [82].

No new data were collected. The study re-analyses material gathered and documented during the doctoral research: electricity consumption from four specimen analysers, sampled at two-minute intervals over 28 consecutive days and yielding more than 10 million timestep records; specimen arrival data from the laboratory information management system; a Sustainable Balanced Scorecard of 29 critical success factors and 52 key performance indicators developed with laboratory and hospital stakeholders; the BCMap linking these data to the scorecard; the LSTM forecasting results for three model configurations; and the record of interviews and validation sessions with seven management and technical roles, including the laboratory director, the assistant laboratory manager, the clinical director and hospital IT managers [29]. Results reported in an earlier article on health governance [30] are cited where they recur; the analyses in Sections 5.3 to 5.5 are new. These operational datasets are not publicly available; the conditions for access are set out in the Data availability statement. Because the thesis is under embargo until 2029, every value used in the analysis is stated in full in this article. The only publicly available dataset used is the UK Government greenhouse gas conversion factor for grid electricity, which is cited where it is applied in Section 5.3 and listed in the Data availability statement.

3.3 Analyses

Four analyses were conducted. A1 maps each component of the instrumented laboratory onto the Grieves physical–virtual–connection model, the ISO 23247 reference domains and the Kritzinger classification of data flows [1, 12, 31]. A2 re-reads the predictive results from the perspective of RQ1, comparing the accuracy and interpretability of the three model configurations. A3 is a retrospective decision-rights audit: every governance decision documented across the five steps of the DBA implementation was identified from the doctoral record, described, and coded against the six criteria of the urban decision-rights function introduced in Section 4.3. A4 is an analytical upscaling exercise that traces how a node-level forecast propagates through the nested BCMap and how the Extrapolation Escalation Rule would act on it. A4 uses only values reported in the doctoral record and introduces no simulated data; it is presented as illustrative. In the terms of the FEDS framework, this is an early

formative evaluation that combines a naturalistic, historical instantiation with analytical walkthroughs; summative naturalistic evaluation in a live city twin is left to future work [83].

3.4 Ethics

The doctoral field study on which this article draws, including the interviews and validation sessions with laboratory and hospital staff, was approved by the University of Liverpool Management School Research Ethics Committee (reference 4406, approved 1 May 2019), and the research was carried out in accordance with the guidelines and conditions of approval of that committee. All participants in the interviews and validation sessions gave informed consent. The study collected only operational data: electricity consumption from current transformers clamped around the analysers' supply cables, and specimen counts from the laboratory information management system. No patient-level records were accessed, and the University of Liverpool's ethics review confirmed that the study was exempt from NHS research and development governance on this basis. Data were transmitted over a dedicated private network rather than the hospital's wireless network [29, 30]. The present article is a secondary analysis of data collected under that approval, for a purpose consistent with the original study; no new data were collected and no further participants were involved.

4 The Decision Twin architecture

4.1 Overview

A Decision Twin is a digital twin whose return path from virtual to physical systems is governed by explicit, auditable decision rights (Fig. 1). It retains the mirror functions of a conventional twin and adds four governance layers, which follow the Predictive Governance Architecture introduced for health governance [30] and are adapted here to city systems (Fig. 2). Layer 1, the institutional layer, holds the legal mandates, decision-rights register and precommitments within which every other layer operates. Layer 2, the data layer, provides semantic interoperability and hosts the nested BCMap that links sensor signals to institutional and city indicators. Layer 3, the intelligence layer, hosts forecasting models, including agentic components, and applies the Extrapolation Escalation Rule. Layer 4, the human interface, is where officials and citizens exercise the authority assigned to them and where drift in that authority is measured.

The key difference from existing twin taxonomies is shown in Fig. 1. Under the Kritzinger classification, a twin matures by automating its return flow [12]. In a Decision Twin, the return flow is differentiated: some actions are executed automatically, some are recommended to an accountable official, and some are referred to deliberative processes. Maturity is measured by how reliably each flow follows its assigned route, not by how much of the loop is automated.

4.2 Nested Business Connectivity Map

The BCMap is a cause-and-effect structure that links operational variables to strategic indicators [29]. The Decision Twin generalises it into a hierarchy of units at four resolutions (Fig. 3): nodes (facilities and

assets), institutions (organisations that own and operate nodes), sectors (energy, health, mobility, water) and the city. At node *n*, a forecasting component F_n predicts the indicator vector one step ahead from a window of *w* past observations:

$$\hat{k}_n(t+1) = F_n(x_n(t-w+1), \ldots, x_n(t)) \quad (1)$$

At HL@RHL, x_n comprised electricity consumption per analyser and specimen arrivals, and the window was 10,080 two-minute timesteps, or 14 days [29]. Each unit *u* above node level composes the forecasts of its children $C(u)$ through mapping matrices $W_{u,c}$, which translate child indicators into parent indicators (for example, kilowatt-hours into a sector energy indicator), and receives cross-sector effects through coupling matrices Γ:

$$\hat{k}_u(t+1) = \Sigma_{c \in C(u)} W_{u,c}\, \hat{k}_c(t+1) + \Sigma_{s' \neq s(u)} \Gamma_{s(u),s'}\, \hat{k}_{s'}(t+1) \quad (2)$$

where $s(u)$ is the sector of unit *u*. The feedforward signal at any level is the gap between the target k^*_u and the forecast:

$$g_u(t+1) = k^*_u(t+1) - \hat{k}_u(t+1) \quad (3)$$

Finally, city-level value is a weighted combination of city indicators:

$$\hat{V}(t+1) = \lambda^\top \hat{k}_{city}(t+1), \text{ with } \lambda_j \geq 0 \text{ and } \Sigma_j \lambda_j = 1 \quad (4)$$

Equations (2) to (4) do not assume that the mappings are linear in practice; they state the composition explicitly so that it can be inspected. Two design choices follow from them. First, the coupling matrices make cross-sector dependencies visible: a reduction in energy use at a hospital is read together with its effect on health-service throughput rather than in isolation (Section 5.3). Second, the weights λ in Eq. (4) are not technical parameters. They express the relative priority a city gives to its objectives, and their setting is a civic–deliberative decision (Section 4.3). The doctoral implementation already used a two-level instance of Eqs. (2) and (4): forecasts for four analysers were summed into a laboratory energy indicator, and laboratory value was computed as a weighted combination of the forecast financial value of tests and forecast operating cost, with weights set from NHS policy and management priorities [29].

4.3 Urban decision-rights function

In health governance, ACI classified a decision type by three criteria: its dominant cognitive capacity *g*, the regulatory oversight obligation it triggers *o*, and the assignment of professional accountability ρ [30]. Urban decisions require three further criteria. Reversibility *r* records whether an action can be undone within the control cycle, undone only at cost, or not undone at all. Equity exposure *e* records whether the consequences fall evenly or concentrate on particular places or groups, particularly those with fewer alternatives [76]. Envelope status *v* records whether the inputs on which a recommendation is based lie inside or outside the support of the model's training data [57, 58]. Together the six criteria operationalise

the tracking and tracing conditions of meaningful human control [70]: reversibility, equity and envelope status test whether a delegated action can still respond to the reasons of those affected, while accountability assignment ensures that outcomes trace to an identifiable person. They also reflect public administration research that treats task complexity, uncertainty and equity as criteria for allocating discretion between officials and AI [66, 67]. Each criterion maps to a minimum domain on an ordered scale I < II < III < IV, and a decision type *d* is assigned to the highest domain that any criterion requires:

$$D(d) = \max\{ D_g(g), D_o(o), D_\rho(\rho), D_r(r), D_e(e), D_v(v) \} \quad (5)$$

The maximum operator embodies a precautionary ratchet: any criterion can raise the domain of a decision, but none can lower it. Table 2 defines the four urban domains. Domains I to III follow the health taxonomy [30]. Domain IV is new. In health governance it covered the relational and therapeutic encounter; in cities it becomes civic–deliberative: decisions about the ends that urban systems should serve, which draw on existential and interpersonal capacities and on democratic legitimacy rather than on expertise [71, 75].

Table 2
Urban decision-rights domains of the Decision Twin

Domain	Authority status	Dominant capacity and triggering criteria	Illustrative urban decisions
I Computational operations	AI may act within the envelope; human oversight required	Logical-mathematical; reversible within the control cycle; low equity exposure; inputs inside the training envelope	Retiming signals within approved plans; shifting building heating, ventilation and air-conditioning loads; reordering routine inspections
II Contextual interpretation	AI recommends; an accountable official decides	Contextual and interpersonal; reversible at cost; differential effects across places; or any Domain I case outside the envelope	Prioritising maintenance across neighbourhoods; responding to an unusual demand surge; rerouting buses around an incident
III Ethical and accountable judgement	AI excluded from the decision; human decision mandatory	Ethical and accountability-bearing; irreversible or legally consequential; high exposure for vulnerable groups	Rationing a utility during scarcity; closing a public service; enforcement action against individuals
IV Civic–deliberative	AI may inform deliberation; citizens and elected bodies decide	Existential and interpersonal; contested ends; long-lived, distributive consequences	Setting the weights λ between city objectives; land-use futures; heritage and displacement trade-offs

4.4 Extrapolation Escalation Rule

Envelope status is the criterion most likely to change from moment to moment, so it is operationalised as a rule. Let $\mathcal{E}$ denote the support of the training data of the model that produced a recommendation, and let δ be a distance from the current input vector to that support, such as a per-feature quantile bound or a Mahalanobis distance. Then

$v(t) = \text{out}$ if $\delta(x(t), \mathcal{E}) > \theta$, otherwise in; $D_v = \text{II}$ if $v(t) = \text{out}$, otherwise I (6)

where θ is a threshold set in Layer 1. The rule states that an action the twin may normally take on its own becomes a recommendation to a human whenever the inputs lie outside the conditions on which the model was trained. It converts the interpolation–extrapolation asymmetry of machine learning [58] into a routing rule. It also answers Bainbridge's irony of automation [56] by design: humans are brought in precisely when the model is least reliable, and the escalation is triggered by a measurable condition rather than by an operator's vigilance.

4.5 Ulysses Protocol

Decision rights that can be relaxed by the system they constrain do not constrain it. The Ulysses Protocol is a set of precommitments, written into Layer 1 and adopted through the city's ordinary democratic procedures, that the agentic layer cannot alter [63, 64]. It has five clauses:

U1 Domain ceiling

agentic components may execute only Domain I actions; Domains III and IV are outside their action space, not merely discouraged.

U2 Non-negotiable escalation

the Extrapolation Escalation Rule cannot be disabled by operators in real time; its threshold θ can be changed only through the amendment procedure in U5.

U3 Judgement floor

a minimum share of Domain II recommendations receives substantive human review, and a ceiling on the Decision-Rights Drift Index triggers an independent audit.

U4 Stop right

a designated public authority can suspend agentic execution at any time, and the twin then reverts to monitoring and recommendation without loss of service.

U5 Amendment friction

changes to U1–U4 require deliberation by an elected body and a cooling-off period before taking effect.

The protocol applies institutionalised distrust [65] to the twin itself, and it answers a familiar argument: that ethical commitments are sufficient because operators have every incentive to behave responsibly. Incentives change under pressure, and the pressure to automate further is constant and local. Ulysses did not rely on his resolve when the Sirens sang; he relied on the rope.

4.6 Decision-Rights Drift Index

The Judgment Preservation Index measures whether individual reviewers diverge from algorithmic recommendations when divergence is warranted [30]. The Decision-Rights Drift Index measures a different quantity: whether the effective authority over a decision type has moved towards the machine relative to its assigned domain. For each decision type j with assigned domain D_j, let $a_j(T)$ be the share of actions of that type in period T that were either executed autonomously or ratified unchanged with review time below a minimum deliberation time τ_{min}. The effective domain is

$$\check{D}_j(T) = I \text{ if } a_j(T) \geq \alpha, \text{ otherwise } D_j \quad (7)$$

and the index, with consequence weights w_j, is

$$DRDI(T) = \Sigma_j \, w_j \, I[\check{D}_j(T) < D_j] \, / \, \Sigma_j \, w_j \quad (8)$$

where $I[\cdot]$ equals 1 when the condition holds and 0 otherwise. DRDI ranges from 0, when every decision type is exercised at its assigned level, to 1, when every non-computational decision type has in effect been delegated to the machine. It is computable from the action and review logs that a Decision Twin already keeps, so it can be monitored continuously as a Layer 4 indicator. The index responds to two findings: oversight requirements tend to become nominal in practice [27], and augmentation tends to slide into automation over time unless managed deliberately [54]. DRDI makes that slide visible before it becomes entrenched. The parameters α and τ_{min} are policy settings and are recorded in the decision-rights register.

5 Evidence from the instrumented node

5.1 The laboratory as a facility-scale twin

HL@RHL was a predominantly automated laboratory: more than 80% of its workload ran on four specimen analysers, so analyser performance governed laboratory performance [29]. Electricity consumption was used as a non-invasive proxy for analyser performance, an approach consistent with research on IoT-based, machine-level energy monitoring in smart factories [84]. LSTM networks are also well established for short-term load forecasting in urban energy systems [85], so the forecasting component transfers naturally to city energy twins. Table 3 maps the instrumentation onto established twin reference models.

Table 3
HL@RHL components mapped to digital twin reference models

Component	HL@RHL implementation [29]	Grieves element	ISO 23247 domain
Physical system	Four specimen analysers on an uninterruptible power supply; laboratory environment	Physical space	Observable physical elements
Sensing	Janitza UMG 604-PRO power-quality analyser; four current transformers clamped around supply cables; CO2 and humidity sensors; specimen arrivals from the TelePath laboratory information system	Connection (physical to virtual)	Device communication
Connectivity and security	Private wireless virtual private network with a dedicated modem and static IP address, segregated from the hospital network	Connection	Device communication
Virtual models	LSTM components per analyser plus response and value predictors, arranged by the BCMap and trained on a high-performance computing cluster	Virtual space	Digital twin
Strategic layer	Sustainable Balanced Scorecard with 29 critical success factors and 52 indicators; DBA gap analysis	Virtual space (services)	User
Return path	Corrective actions decided and executed by laboratory managers	Connection (virtual to physical), human-mediated	User

Under the Kritzinger classification the node is a digital shadow: data flowed automatically from physical to virtual, but the return flow passed through managers [12]. Read through the Decision Twin lens, that is not a deficiency. Every return flow at HL@RHL involved contextual or accountable judgement (Section 5.4), so routing it through people was the correct allocation. The segregated network also shows that the security of a twin's connection layer can be designed in from the start [38].

5.2 Predictive feedforward control at node level

The doctoral study compared three configurations of the BCMap components [29]. Scenario A was the most disaggregated: a separate LSTM component forecast the consumption of each analyser, these forecasts were linked to specimen arrivals according to the BCMap, and laboratory value was computed as a weighted combination in which the weights were set from NHS policy and management priorities. Scenario B used two components, one for total consumption and one for specimen response. Scenario C pooled all data into a single network. Each network had four hidden layers of 50 units, used the Adam optimiser and dropout of 0.2, and was trained on 21 days and tested on 7 [29, 30].

Scenario A forecast most accurately (RMSE 0.00436), followed by B (0.0056) and C (0.0090); the pooled model's error was about twice that of the human-weighted, disaggregated model (Fig. 4). In validation sessions, laboratory and hospital managers unanimously preferred Scenario A for its balance, clarity and practicability, and noted that it combined human and machine forecasts into a form of collective intelligence. They judged Scenario C comprehensive but warned that pooling all data risked losing crucial facts in the mass [29]. Management also estimated that deviations identified one day ahead would otherwise have gone undetected for two to four weeks under the retrospective scorecard [30].

These results bear directly on RQ1. Pre-emptive, indicator-linked control was feasible at node level with operational data alone. More importantly for city twins, the configuration that was easiest to interpret, and that embedded human value judgements explicitly, was also the most accurate. This is consistent with Rudin's argument against black boxes in high-stakes settings [60] and with the complementarity principle of ACI: the gain came from structuring the model around accountable human priorities, not from removing them.

5.3 A cross-sector reading of a day-ahead forecast

On day 21 the model forecast the indicators for day 22 [29]. Table 4 compares the forecasts with the reference values reported for that day in the doctoral record. Re-examination of the raw meter records for this article showed that the electricity reference value is the metered total for day 22 and that it was recorded as the sum of two-minute current readings rather than in kilowatt-hours. Table 4 therefore converts the electricity figures to kWh at 225 V and a power factor of 0.98, the values used in the doctoral study [29]. The conversion does not change the relative gaps.

Table 4
Day-ahead forecast against reference values at HL@RHL (day 22)

Indicator	Reference value [29]	Day-ahead forecast	Gap
Electricity consumption, all analysers (kWh)	33.4	33.2	−0.24 (−0.7%)
Emissions (kg CO2e)	9.46	9.39	−0.07 (−0.7%)
Specimens processed	1,095	1,044	−51 (− 4.7%)

The emissions figures follow from the electricity figures at the UK Government's 2018 greenhouse gas conversion factor for grid electricity, 0.28307 kg CO2e per kWh [86], which the doctoral study also used [29]. The coupling matters at city scale because health care is a material source of greenhouse gas emissions, nationally and globally [87, 88], while also being an essential service whose throughput cannot be traded away for carbon savings. Read in isolation, the energy and carbon gaps are good news: the laboratory is forecast to consume less and emit less than its reference values indicate. Read through the BCMap, which links analyser energy to specimen throughput, they are the signature of under-utilisation: the laboratory is forecast to process 51 fewer specimens than its reference value. An energy-

sector twin that aggregated hospital consumption without the coupling term Γ in Eq. (2) would record a saving and could reward it; a Decision Twin records a probable service shortfall in the health sector and an energy reduction that is its consequence. Because the electricity and specimen series in the doctoral record were collected in different periods, this reading illustrates how the coupling term operates rather than estimating its size.

This small example illustrates the cost of examining urban sub-systems in silos, a limitation that socio-technical research on city twins has emphasised [15]. It also shows why the coupling matrices belong in the governance layer rather than being left implicit in model code. Whether a drop in energy use is efficiency or under-provision is not a statistical fact but an interpretation, and it determines what the city should do next.

5.4 Retrospective decision-rights audit

Table 5 lists the twelve governance decisions documented across the five DBA steps at HL@RHL and the domain each would be assigned under Eq. (5).

Table 5
Decision-rights audit of the DBA implementation at HL@RHL

#	Decision (DBA step)	Taken by	Criteria raising the domain	Domain
1	Map laboratory strategy and select critical success factors (Step 1)	Laboratory and hospital managers, through interviews	Accountability for strategic priorities	III
2	Define 52 indicators and their targets (Step 1)	Managers with the researcher	Contextual knowledge of laboratory work	II
3	Adopt energy as a proxy for analyser performance and specify sensors (Step 2)	Researcher, with laboratory staff	Contextual judgement about what the data represent	II
4	Route data over a private network rather than the hospital network (Step 2)	Hospital IT managers	Legal and security accountability	III
5	Clean and synthesise sensor and laboratory data (Step 2)	Automated pipeline, checked by the researcher	None beyond validation	I
6	Specify cause-and-effect links in the BCMap (Step 3)	Researcher and managers, iteratively	Contextual interpretation of causal structure	II
7	Train models and forecast indicators (Step 4)	LSTM components	None beyond validation	I
8	Choose model configurations to compare (Step 4)	Researcher	Contextual judgement on aggregation	II
9	Set value weights between throughput and cost, $\lambda = 0.7$ (Step 4)	Researcher, from NHS policy and management priorities	Value priority with accountability	III
10	Detect day-ahead variances against targets (Step 5)	Model output	None beyond validation	I
11	Select corrective actions, such as rescheduling energy-intensive work or cross-training staff (Step 5)	Laboratory managers	Context; effects on staff and service	II
12	Adopt Scenario A for future decision-making (validation)	Laboratory and hospital management	Institutional accountability	III

Three of the twelve decisions were computational (Domain I); five required contextual interpretation (Domain II); four carried accountability that only managers could bear (Domain III). No decision fell into Domain IV. Two findings follow. First, even in a highly automated facility where models forecast with high accuracy, three quarters of the decisions that made the system work depended on human judgement. The model's accuracy was a product of that judgement, since decisions 1, 6 and 9 defined what the

model forecast and how its outputs were valued. Second, the absence of Domain IV decisions is itself informative. Within a single institution, the ends are given by policy. At city scale they are not, and the weights λ in Eq. (4) become a civic choice. The civic–deliberative domain is therefore not an ethical embellishment; it is the decision class that appears when a twin crosses the boundary of a single organisation.

5.5 Illustrative upscaling and escalation

Figure 5 traces the day-22 forecast through the Decision Twin. At node level, the variance in specimens is a Domain I detection (decision 10). Propagated through Eq. (2), it reaches two sector indicators: health-service throughput, where it appears as a shortfall, and city energy demand, where it appears as a saving. Because the coupling term links the two, the response is no longer computational: whether to restore throughput, and at what energy cost, depends on context and on effects on patients and staff. Under Eq. (5), the response is therefore a Domain II decision for the laboratory's accountable managers, with the forecast as evidence.

Suppose instead that specimen arrivals on day 21 had exceeded every value observed during the 21-day training window, as might occur during a major incident or an epidemic surge. The Extrapolation Escalation Rule would mark the input as out of envelope, and even the detection step would be demoted to Domain II: the forecast would be presented as low-confidence evidence rather than a reliable signal, and no automated rescheduling would be permitted. The doctoral fieldwork was itself conducted under pandemic restrictions that limited access to the laboratory [29], a reminder that such conditions are not hypothetical. No new numbers are introduced in this walkthrough; it shows how the rules of Section 4 act on the reported forecast.

6 Discussion

6.1 Theoretical contributions

The study contributes to four bodies of theory. For digital twin research, it reframes the return path. Existing classifications rank twins by how much of the loop they close automatically [12]. The Decision Twin replaces a single maturity scale with a routing problem: each return flow is assigned to a domain, and a twin is mature when flows reliably follow their assigned routes. This makes explicit what socio-technical studies have argued, that the value of a city twin depends on how it is embedded in institutions [15], and gives architectural form to the view of twins as steering techniques [40]. It also departs from maturity models that place autonomous decision-making and implementation at the top of the scale [39]: in a Decision Twin, autonomy is a bounded property of Domain I decisions, not the goal of the whole system.

For management control, the study extends predictive feedforward control from a single organisation to a system of organisations. The nested BCMap shows how forecasts at facility level can be composed into the indicators of institutions, sectors and the city, and how cross-sector coupling changes the

meaning of a variance (Section 5.3). This addresses the cybernetic requirement that a controller match the variety of the system it regulates [44] without centralising decisions, since every level keeps its own gap signal and its own decision rights.

For human–AI boundary theory, the study extends ACI from health governance [30] to cities. Reversibility, equity exposure and envelope status join the original criteria, the domains are combined by a precautionary maximum, and Domain IV is respecified as civic–deliberative. The Extrapolation Escalation Rule turns a known statistical weakness of machine learning [57, 58] into a governance trigger, so that people are brought in when models are least reliable rather than when operators happen to be vigilant [56].

For institutional theory, the study applies precommitment to algorithmic governance. North and Williamson explain why rules and governance structures matter when actors are boundedly rational [61, 62]; Elster explains why actors sometimes bind themselves in advance [63, 64]. The Ulysses Protocol combines both: a city binds its own twin, and itself, before the pressures that would erode human authority arrive. It complements principle-based AI ethics, which lacks enforcement [23], with enforceable architecture, and it answers the critique that human-oversight requirements become nominal in practice [27] by specifying where oversight is required and measuring whether it still occurs.

6.2 The twin is not the asset

A strategic implication follows from the resource-based view. Twin platforms, sensors and forecasting models are becoming standardised, open and widely available. Resources that are widely available and imitable cannot be a lasting source of advantage [49]; the value of a technology is captured instead by the complementary assets that surround it [50]. For cities, the analogue of competitive advantage is public value. The evidence from HL@RHL suggests where that value lies: the most accurate configuration was the one structured around explicit human priorities, and three quarters of the decisions that made the system work were human. The governance configuration, comprising the decision-rights register, the precommitments, the trained officials and the trust of citizens, is specific to each city, accumulates slowly and cannot be purchased with the platform. This is consistent with reviews that rank practical value among the most severe unresolved challenges of urban twins and that list governance, organisational and social issues among them [10, 11] and with research showing that smart urban governance depends on socio-technical collaboration rather than technology [74]. Seify called the corresponding health-sector capability the Governance Integration Layer [30]; the urban equivalent may be termed a Civic Governance Integration Layer. Three propositions follow for future testing:

P1

The public value that cities obtain from digital twins will depend increasingly on governance complements and decreasingly on twin technology, as twin platforms commoditise.

P2

Cities that allocate decision rights before deploying agentic components will show lower decision-rights drift over time than cities that allocate them after deployment.

P3

Twins that model cross-sector coupling explicitly will produce fewer interventions that improve one sector's indicators at another sector's expense than twins that aggregate sector indicators independently.

6.3 Participation and civic legitimacy

Domain IV gives participatory governance a defined place in the architecture. The Collection calls for human-centric urban intelligence and participatory decision-making; the Decision Twin specifies where participation is required rather than merely desirable. Setting the weights λ between city objectives, deciding land-use futures and resolving heritage and displacement trade-offs are decisions about ends. They belong at the upper rungs of Arnstein's ladder [72] and correspond to Rahwan's society-in-the-loop, in which the values encoded in algorithmic systems are negotiated with the public [75]. Smart-city programmes have mostly cast citizens as consumers or as subjects of stewardship rather than as holders of decision rights [73]; Domain IV moves them up by design, and participatory twins such as those in Herrenberg and in citizen-feedback platforms show that the interfaces already exist [8, 37]. Twins can support this domain by simulating the consequences of alternative weightings, which is a Domain I contribution to a Domain IV decision. In Maslow's later terms, the question these decisions answer is not what the city can become but what it exists for [77, 78]. If the most important decisions about urban AI are to be democratic, they cannot be settled by technology suppliers or optimisation routines alone.

6.4 A maturity pathway for city twins

Figure 6 sets out a five-stage maturity pathway. Stage 0 is the mirror twin, which replicates state. Stage 1 adds predictive feedforward: forecasts are linked to targets and gaps are reported before deviations occur. Stage 2 adds governance: a decision-rights register, the Ulysses Protocol and DRDI monitoring. Stage 3 permits bounded agency: Domain I actions are executed automatically inside the envelope. Stage 4 adds civic integration: nested BCMaps span sectors and Domain IV decisions are supported by deliberative interfaces. The order matters. Under this pathway a city does not grant agency (Stage 3) before it has built governance (Stage 2), and the highest stage is civic integration rather than full autonomy, in contrast to automation-led maturity models [39].

The horizons in Fig. 6 are the author's foresight, offered as a structured reading of plausible developments rather than a forecast. By 2027, organisational AI risk frameworks and baseline AI literacy are likely to become common, which supports Stages 1 and 2. By 2028, hybrid human–AI work, AI audit roles and the orchestration of multiple agents may become normal practice. By 2029, agentic systems capable of managing end-to-end processes are likely to be available to city operators, which makes

Stage 3 technically possible; the pathway argues that it should be adopted only where Stage 2 is in place. From 2030 onwards, as AI becomes ubiquitous, the differentiating capabilities of cities are likely to be systems thinking, judgement under ambiguity and civic deliberation about purpose, the capabilities that Stage 4 institutionalises.

6.5 Implications for practice and policy

Four implications follow for city governments, regulators and suppliers. First, procurement should require suppliers of twin and urban-AI systems to classify every proposed decision type against the four domains before contract award, so that the allocation of decision rights is a specification rather than an emergent property of deployment. The AI Act's oversight requirements for critical-infrastructure systems [28] can be met more precisely in this way, because the domain assignment states where oversight applies and in what form. Table 6 shows how four entries in such a decision-rights register would be completed under Eq. (5).

Table 6
Illustrative entries in a city decision-rights register

Decision type	Criteria (g; o; ρ; r; e)	Envelope rule	Domain	Ulysses clause
Retime signals at a junction within an approved signal plan	Logical-mathematical; high-risk where used as a safety component in road traffic; validation only; reversible within the cycle; low exposure	Demote to II when traffic inputs leave the training envelope	I	U2
Shift heating and cooling loads across a hospital estate during a heatwave	Contextual; high-risk where a safety component in heating or electricity supply; shared between estates and clinical leads; reversible at cost; differential exposure for patients	Applies; recommendation flagged as low-confidence when out of envelope	II	U3
Reduce water pressure across districts during a drought	Ethical and accountable; high-risk where a safety component in water supply; exclusive with legal consequence; partly irreversible; high exposure for vulnerable households	Not applicable: AI outside the action space	III	U1
Set the weights λ between carbon, service access and cost in the city twin	Civic–existential; policy rather than system output; elected body; long-lived; distributive	Not applicable	IV	U1, U5

Second, cities should adopt Ulysses clauses in the governance charters of their twins, with the amendment friction of clause U5. A stop right and a non-negotiable escalation rule are inexpensive at design time and very costly to retrofit.

Third, cities should plan for workforce change. Automation displaces some tasks and creates others [89]; the World Economic Forum's 2020 survey of employers, for example, paired an estimated 85 million displaced roles with 97 million new ones by 2025 [90]. In a Decision Twin, routine monitoring tasks move to Domain I, while new roles emerge around the domains that remain human: decision stewards who own the decision-rights register, internal algorithmic auditors [91], and facilitators of civic deliberation. Training should distinguish data literacy, which concerns the quality, bias and meaning of data, from AI literacy, which concerns model behaviour, failure modes and the points at which human judgement must intervene. Both are needed for officials to exercise Domain II and III authority with confidence rather than by deference [24, 92], and to hold algorithmic systems to account [68].

Fourth, security should be designed into the connection layer from the start. The segregated network at HL@RHL is a small instance of a principle that matters more as twins gain agency: a compromised twin that can act is more dangerous than one that can only observe [38].

7 Limitations and future research

The study has several limitations. The empirical demonstration rests on one node, one laboratory and 28 days of data collected during doctoral research [29]; it shows that the architecture can be instantiated, not that it generalises. The upscaling in Section 5.5 is analytical and introduces no new data, so the behaviour of the nested BCMap across many nodes and sectors remains to be tested. In the original modelling, hyperparameters were selected against test-set performance rather than a separate validation split [30], which may flatter the reported accuracy, although the comparison between configurations was made under the same conditions. The retrospective audit was coded by one researcher who also conducted the original study; independent coding with an inter-rater agreement statistic would strengthen it. The Ulysses Protocol and the DRDI are proposed and specified but not yet validated in operation, and the parameters α, τ_{min} and θ require empirical calibration. Finally, Gardner's theory [48] is not universally accepted in cognitive science; the decision-rights function reduces dependence on it by combining cognitive capacity with accountability, reversibility, equity and envelope criteria, but alternative theoretical bases deserve systematic comparison.

These limitations define an agenda. Following the FEDS framework [83], the next step is a summative, naturalistic evaluation in a live city twin. A multi-node pilot spanning at least two sectors, such as a hospital estate and a district energy network, would test the nested BCMap and proposition P3. A prospective deployment of the DRDI alongside the Judgment Preservation Index would show whether authority drift can be detected before it becomes entrenched. System-dynamics and agent-based simulation, recommended in the doctoral research as a response to data restrictions [29], could explore how different Ulysses clauses perform under stress. Comparative studies of cities that adopt decision-rights registers early and late would test proposition P2.

8 Conclusion

City digital twins are acquiring the ability to act. This article has argued that the critical question is therefore no longer how faithfully a twin mirrors the city, but how the decisions taken through it are governed. The Decision Twin answers the three research questions as follows. For RQ1, predictive feedforward control linked to institutional indicators is feasible with operational data alone: at the instrumented laboratory, a human-weighted disaggregated configuration forecast one day ahead with an RMSE of 0.00436, about half the error of a pooled model. For RQ2, decision rights can be allocated by a precautionary function of six criteria, which places computational operations, contextual interpretation, accountable judgement and civic deliberation in four domains and demotes any machine action to human decision when its inputs leave the training envelope. For RQ3, the nested BCMap composes node-level control to city scale, the Ulysses Protocol holds the allocation of rights in place, and the Decision-Rights Drift Index makes the erosion of human authority measurable.

The evidence from a single, highly automated facility is modest in scale but clear in direction: only three of twelve governance decisions were computational, and a reduction in energy use that a siloed twin would have counted as a saving signalled a shortfall in health-service throughput. As twin platforms commoditise, the value cities obtain from them will depend less on the twin and more on the governance around it. AI should serve as the city's assistant; people, through accountable institutions and civic deliberation, should remain responsible for its direction.

Declarations

Acknowledgements

The author thanks the management and staff of the Haematology and Coagulation Laboratory at the Royal Liverpool Hospital for their participation in the doctoral research on which this article draws, the doctoral supervisors at the University of Liverpool, and the University of Liverpool's Advanced Research Computing service for access to the Barkla high-performance computing cluster.

Author contributions

M.S. conceived the study, developed the Decision Twin architecture, conducted the original doctoral fieldwork and analysis, performed the secondary analyses, and wrote and revised the manuscript.

Funding

This article received no specific funding. The doctoral research on which it draws was supported by the European Regional Development Fund through the Eco-I North West programme at the University of Liverpool, with industry support from CSols Ltd.

Data availability

The datasets generated during and/or analysed during the current study are not publicly available due to NHS data governance restrictions on the operational records of the host hospital laboratory but are

available from the corresponding author on reasonable request. The publicly available reference dataset used in this study is the UK Government greenhouse gas conversion factor for grid electricity, published by the Department for Business, Energy and Industrial Strategy in Greenhouse gas reporting: conversion factors 2018 and available at https://www.gov.uk/government/publications/greenhouse-gas-reporting-conversion-factors-2018 [86].

Code availability

The Python code used for the original LSTM modelling is reported in the appendix of the author's doctoral thesis [29], which is under embargo until 2029; it is available from the corresponding author on reasonable request.

Ethics approval and consent to participate The doctoral field study on which this article draws was approved by the University of Liverpool Management School Research Ethics Committee (reference 4406, approved 1 May 2019), and the research was carried out in accordance with the guidelines of the University of Liverpool Management School Research Ethics Committee. All participants in the interviews and validation sessions gave informed consent to participate. The study collected only operational, non-patient data, and the University of Liverpool's ethics review confirmed that it was exempt from NHS research and development governance on this basis. This article is a secondary analysis of data collected under that approval; no new data were collected.

Consent for publication Not applicable.

Competing interests The author is the founder of VisionXY7 Ltd, an AI consultancy. No VisionXY7 data, products or funding were used in this study. The author declares no other competing interests.

Use of AI-assisted tools The research idea, design, data, analyses, arguments and overall structure of this article are the author's own and derive from his doctoral research. Claude (Anthropic), used between October and November 2026, helped prepare a first draft of the text from the author's own material and was used for language editing. The author reviewed, revised and verified all content and takes full responsibility for the manuscript.

References

1. Grieves M, Vickers J. Digital twin: mitigating unpredictable, undesirable emergent behavior in complex systems. In: Kahlen FJ, Flumerfelt S, Alves A, editors. Transdisciplinary perspectives on complex systems. Cham: Springer; 2017. p. 85–113. https://doi.org/10.1007/978-3-319-38756-7_4
2. Batty M. Digital twins. Environ Plan B Urban Anal City Sci. 2018;45(5):817–20. https://doi.org/10.1177/2399808318796416
3. Ketzler B, Naserentin V, Latino F, Zangelidis C, Thuvander L, Logg A. Digital twins for cities: a state of the art review. Built Environ. 2020;46(4):547–73. https://doi.org/10.2148/benv.46.4.547

4. Shahat E, Hyun CT, Yeom C. City digital twin potentials: a review and research agenda. Sustainability. 2021;13(6):3386. https://doi.org/10.3390/su13063386
5. Deng T, Zhang K, Shen ZJM. A systematic review of a digital twin city: a new pattern of urban governance toward smart cities. J Manag Sci Eng. 2021;6(2):125–34. https://doi.org/10.1016/j.jmse.2021.03.003
6. Farsi M, Daneshkhah A, Hosseinian-Far A, Jahankhani H, editors. Digital twin technologies and smart cities. Cham: Springer; 2020. https://doi.org/10.1007/978-3-030-18732-3
7. Schrotter G, Hürzeler C. The digital twin of the city of Zurich for urban planning. PFG J Photogramm Remote Sens Geoinf Sci. 2020;88(1):99–112. https://doi.org/10.1007/s41064-020-00092-2
8. Dembski F, Wössner U, Letzgus M, Ruddat M, Yamu C. Urban digital twins for smart cities and citizens: the case study of Herrenberg, Germany. Sustainability. 2020;12(6):2307. https://doi.org/10.3390/su12062307
9. Ferré-Bigorra J, Casals M, Gangolells M. The adoption of urban digital twins. Cities. 2022;131:103905. https://doi.org/10.1016/j.cities.2022.103905
10. Lei B, Janssen P, Stoter J, Biljecki F. Challenges of urban digital twins: a systematic review and a Delphi expert survey. Autom Constr. 2023;147:104716. https://doi.org/10.1016/j.autcon.2022.104716
11. Weil C, Bibri SE, Longchamp R, Golay F, Alahi A. Urban digital twin challenges: a systematic review and perspectives for sustainable smart cities. Sustain Cities Soc. 2023;99:104862. https://doi.org/10.1016/j.scs.2023.104862
12. Kritzinger W, Karner M, Traar G, Henjes J, Sihn W. Digital twin in manufacturing: a categorical literature review and classification. IFAC-PapersOnLine. 2018;51(11):1016–22. https://doi.org/10.1016/j.ifacol.2018.08.474
13. Kitchin R. The real-time city? Big data and smart urbanism. GeoJournal. 2014;79(1):1–14. https://doi.org/10.1007/s10708-013-9516-8
14. Kitchin R, Lauriault TP, McArdle G. Knowing and governing cities through urban indicators, city benchmarking and real-time dashboards. Reg Stud Reg Sci. 2015;2(1):6–28. https://doi.org/10.1080/21681376.2014.983149
15. Nochta T, Wan L, Schooling JM, Parlikad AK. A socio-technical perspective on urban analytics: the case of city-scale digital twins. J Urban Technol. 2021;28(1–2):263–87. https://doi.org/10.1080/10630732.2020.1798177
16. Chan A, Salganik R, Markelius A, Pang C, Rajkumar N, Krasheninnikov D, et al. Harms from increasingly agentic algorithmic systems. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. New York: ACM; 2023. p. 651–66. https://doi.org/10.1145/3593013.3594033
17. Cugurullo F. Urban artificial intelligence: from automation to autonomy in the smart city. Front Sustain Cities. 2020;2:38. https://doi.org/10.3389/frsc.2020.00038

18. Cugurullo F, Caprotti F, Cook M, Karvonen A, McGuirk P, Marvin S. The rise of AI urbanism in post-smart cities: a critical commentary on urban artificial intelligence. Urban Stud. 2024;61(6):1168–82. https://doi.org/10.1177/00420980231203386
19. Harari YN. Nexus: a brief history of information networks from the Stone Age to AI. New York: Random House; 2024.
20. Simons R. Levers of control: how managers use innovative control systems to drive strategic renewal. Boston: Harvard Business School Press; 1995.
21. Kaplan RS, Norton DP. The balanced scorecard: translating strategy into action. Boston: Harvard Business School Press; 1996.
22. Otley D. Performance management: a framework for management control systems research. Manag Account Res. 1999;10(4):363–82. https://doi.org/10.1006/mare.1999.0115
23. Mittelstadt B. Principles alone cannot guarantee ethical AI. Nat Mach Intell. 2019;1(11):501–7. https://doi.org/10.1038/s42256-019-0114-4
24. Parasuraman R, Riley V. Humans and automation: use, misuse, disuse, abuse. Hum Factors. 1997;39(2):230–53. https://doi.org/10.1518/001872097778543886
25. Skitka LJ, Mosier KL, Burdick M. Does automation bias decision-making? Int J Hum Comput Stud. 1999;51(5):991–1006. https://doi.org/10.1006/ijhc.1999.0252
26. Alon-Barkat S, Busuioc M. Human–AI interactions in public sector decision making: “automation bias” and “selective adherence” to algorithmic advice. J Public Adm Res Theory. 2023;33(1):153–69. https://doi.org/10.1093/jopart/muac007
27. Green B. The flaws of policies requiring human oversight of government algorithms. Comput Law Secur Rev. 2022;45:105681. https://doi.org/10.1016/j.clsr.2022.105681
28. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Off J Eur Union. 2024;L 2024/1689. http://data.europa.eu/eli/reg/2024/1689/oj
29. Seify M. Development of an advanced artificial intelligence-based model (deep business analytics) for managing and improving control and decision making in modern organisations: application in a hospital clinical laboratory [PhD thesis]. Liverpool: University of Liverpool; 2025. https://doi.org/10.17638/03194278
30. Seify M. Deep business analytics and artificial complementary intelligence: a pre-emptive control framework for human–AI integration in digital health commodity networks. Commodities. 2026;5(3):20. https://doi.org/10.3390/commodities5030020
31. International Organization for Standardization. ISO 23247-2:2021 Automation systems and integration—digital twin framework for manufacturing—part 2: reference architecture. Geneva: ISO; 2021.
32. International Organization for Standardization, International Electrotechnical Commission. ISO/IEC 30173:2023 Digital twin—concepts and terminology. Geneva: ISO; 2023.

33. Batty M. Digital twins in city planning. Nat Comput Sci. 2024;4(3):192–9. https://doi.org/10.1038/s43588-024-00606-7
34. Lu Q, Parlikad AK, Woodall P, Don Ranasinghe G, Xie X, Liang Z, et al. Developing a digital twin at building and city levels: case study of West Cambridge campus. J Manag Eng. 2020;36(3):05020004. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000763
35. Bolton A, Butler L, Dabson I, Enzer M, Evans M, Fenemore T, et al. Gemini principles. Cambridge: Centre for Digital Built Britain; 2018. https://doi.org/10.17863/CAM.32260
36. Charitonidou M. Urban scale digital twins in data-driven society: challenging digital universalism in urban planning decision-making. Int J Archit Comput. 2022;20(2):238–53. https://doi.org/10.1177/14780771211070005
37. White G, Zink A, Codecà L, Clarke S. A digital twin smart city for citizen feedback. Cities. 2021;110:103064. https://doi.org/10.1016/j.cities.2020.103064
38. Alcaraz C, Lopez J. Digital twin: a comprehensive survey of security threats. IEEE Commun Surv Tutor. 2022;24(3):1475–503. https://doi.org/10.1109/COMST.2022.3171465
39. Haraguchi M, Funahashi T, Biljecki F. Assessing governance implications of city digital twin technology: a maturity model approach. Technol Forecast Soc Change. 2024;204:123409. https://doi.org/10.1016/j.techfore.2024.123409
40. Korenhof P, Blok V, Kloppenburg S. Steering representations—towards a critical understanding of digital twins. Philos Technol. 2021;34(4):1751–73. https://doi.org/10.1007/s13347-021-00484-1
41. Malmi T, Brown DA. Management control systems as a package—opportunities, challenges and research directions. Manag Account Res. 2008;19(4):287–300. https://doi.org/10.1016/j.mar.2008.09.003
42. Ferreira A, Otley D. The design and use of performance management systems: an extended framework for analysis. Manag Account Res. 2009;20(4):263–82. https://doi.org/10.1016/j.mar.2009.07.003
43. Figge F, Hahn T, Schaltegger S, Wagner M. The sustainability balanced scorecard—linking sustainability management to business strategy. Bus Strategy Environ. 2002;11(5):269–84. https://doi.org/10.1002/bse.339
44. Ashby WR. An introduction to cybernetics. London: Chapman & Hall; 1956.
45. International Organization for Standardization. ISO 37122:2019 Sustainable cities and communities —indicators for smart cities. Geneva: ISO; 2019.
46. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80. https://doi.org/10.1162/neco.1997.9.8.1735
47. Gardner H. Frames of mind: the theory of multiple intelligences. New York: Basic Books; 1983.
48. Gardner H. Intelligence reframed: multiple intelligences for the 21st century. New York: Basic Books; 1999.

49. Barney J. Firm resources and sustained competitive advantage. J Manag. 1991;17(1):99–120. https://doi.org/10.1177/014920639101700108
50. Teece DJ. Profiting from technological innovation: implications for integration, collaboration, licensing and public policy. Res Policy. 1986;15(6):285–305. https://doi.org/10.1016/0048-7333(86)90027-2
51. Malone TW. Superminds: the surprising power of people and computers thinking together. New York: Little, Brown; 2018.
52. Dellermann D, Ebel P, Söllner M, Leimeister JM. Hybrid intelligence. Bus Inf Syst Eng. 2019;61(5):637–43. https://doi.org/10.1007/s12599-019-00595-2
53. Jarrahi MH. Artificial intelligence and the future of work: human-AI symbiosis in organizational decision making. Bus Horiz. 2018;61(4):577–86. https://doi.org/10.1016/j.bushor.2018.03.007
54. Raisch S, Krakowski S. Artificial intelligence and management: the automation–augmentation paradox. Acad Manage Rev. 2021;46(1):192–210. https://doi.org/10.5465/amr.2018.0072
55. Parasuraman R, Sheridan TB, Wickens CD. A model for types and levels of human interaction with automation. IEEE Trans Syst Man Cybern A Syst Hum. 2000;30(3):286–97. https://doi.org/10.1109/3468.844354
56. Bainbridge L. Ironies of automation. Automatica. 1983;19(6):775–9. https://doi.org/10.1016/0005-1098(83)90046-8
57. Quiñonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND, editors. Dataset shift in machine learning. Cambridge, MA: MIT Press; 2009.
58. Xu K, Zhang M, Li J, Du SS, Kawarabayashi K, Jegelka S. How neural networks extrapolate: from feedforward to graph neural networks. In: International Conference on Learning Representations; 2021. https://doi.org/10.48550/arXiv.2009.11848
59. Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, et al. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In: Advances in Neural Information Processing Systems 32 (NeurIPS 2019). 2019. https://doi.org/10.48550/arXiv.1906.02530
60. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206–15. https://doi.org/10.1038/s42256-019-0048-x
61. North DC. Institutions, institutional change and economic performance. Cambridge: Cambridge University Press; 1990.
62. Williamson OE. The economic institutions of capitalism: firms, markets, relational contracting. New York: Free Press; 1985.
63. Elster J. Ulysses and the sirens: studies in rationality and irrationality. Cambridge: Cambridge University Press; 1979.
64. Elster J. Ulysses unbound: studies in rationality, precommitment, and constraints. Cambridge: Cambridge University Press; 2000.

65. Laux J. Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design of AI governance under the European Union AI Act. AI Soc. 2024;39(6):2853–66. https://doi.org/10.1007/s00146-023-01777-z
66. Bullock JB. Artificial intelligence, discretion, and bureaucracy. Am Rev Public Adm. 2019;49(7):751–61. https://doi.org/10.1177/0275074019856123
67. Young MM, Bullock JB, Lecy JD. Artificial discretion as a tool of governance: a framework for understanding the impact of artificial intelligence on public administration. Perspect Public Manag Gov. 2019;2(4):301–13. https://doi.org/10.1093/ppmgov/gvz014
68. Busuioc M. Accountable artificial intelligence: holding algorithms to account. Public Adm Rev. 2021;81(5):825–36. https://doi.org/10.1111/puar.13293
69. Danaher J. The threat of algocracy: reality, resistance and accommodation. Philos Technol. 2016;29(3):245–68. https://doi.org/10.1007/s13347-015-0211-1
70. Santoni de Sio F, van den Hoven J. Meaningful human control over autonomous systems: a philosophical account. Front Robot AI. 2018;5:15. https://doi.org/10.3389/frobt.2018.00015
71. Rittel HWJ, Webber MM. Dilemmas in a general theory of planning. Policy Sci. 1973;4(2):155–69. https://doi.org/10.1007/BF01405730
72. Arnstein SR. A ladder of citizen participation. J Am Inst Plann. 1969;35(4):216–24. https://doi.org/10.1080/01944366908977225
73. Cardullo P, Kitchin R. Being a 'citizen' in the smart city: up and down the scaffold of smart citizen participation in Dublin, Ireland. GeoJournal. 2019;84(1):1–13. https://doi.org/10.1007/s10708-018-9845-8
74. Meijer A, Bolívar MPR. Governing the smart city: a review of the literature on smart urban governance. Int Rev Adm Sci. 2016;82(2):392–408. https://doi.org/10.1177/0020852314564308
75. Rahwan I. Society-in-the-loop: programming the algorithmic social contract. Ethics Inf Technol. 2018;20(1):5–14. https://doi.org/10.1007/s10676-017-9430-8
76. Eubanks V. Automating inequality: how high-tech tools profile, police, and punish the poor. New York: St. Martin's Press; 2018.
77. Maslow AH. The farther reaches of human nature. New York: Viking; 1971.
78. Koltko-Rivera ME. Rediscovering the later version of Maslow's hierarchy of needs: self-transcendence and opportunities for theory, research, and unification. Rev Gen Psychol. 2006;10(4):302–17. https://doi.org/10.1037/1089-2680.10.4.302
79. Hevner AR, March ST, Park J, Ram S. Design science in information systems research. MIS Q. 2004;28(1):75–105. https://doi.org/10.2307/25148625
80. Peffers K, Tuunanen T, Rothenberger MA, Chatterjee S. A design science research methodology for information systems research. J Manag Inf Syst. 2007;24(3):45–77. https://doi.org/10.2753/MIS0742-1222240302

81. Gregor S, Hevner AR. Positioning and presenting design science research for maximum impact. MIS Q. 2013;37(2):337–55. https://doi.org/10.25300/MISQ/2013/37.2.01
82. Yin RK. Case study research and applications: design and methods. 6th ed. Thousand Oaks, CA: SAGE; 2018.
83. Venable J, Pries-Heje J, Baskerville R. FEDS: a framework for evaluation in design science research. Eur J Inf Syst. 2016;25(1):77–89. https://doi.org/10.1057/ejis.2014.36
84. Shrouf F, Ordieres-Meré J, Miragliotta G. Smart factories in Industry 4.0: a review of the concept and of energy management approached in production based on the Internet of Things paradigm. In: 2014 IEEE International Conference on Industrial Engineering and Engineering Management. Piscataway, NJ: IEEE; 2014. p. 697–701. https://doi.org/10.1109/IEEM.2014.7058728
85. Kong W, Dong ZY, Jia Y, Hill DJ, Xu Y, Zhang Y. Short-term residential load forecasting based on LSTM recurrent neural network. IEEE Trans Smart Grid. 2019;10(1):841–51. https://doi.org/10.1109/TSG.2017.2753802
86. Department for Business, Energy and Industrial Strategy. Greenhouse gas reporting: conversion factors 2018 [dataset]. London: Department for Business, Energy and Industrial Strategy; 2018. https://www.gov.uk/government/publications/greenhouse-gas-reporting-conversion-factors-2018
87. Lenzen M, Malik A, Li M, Fry J, Weisz H, Pichler PP, et al. The environmental footprint of health care: a global assessment. Lancet Planet Health. 2020;4(7):e271–9. https://doi.org/10.1016/S2542-5196(20)30121-2
88. Tennison I, Roschnik S, Ashby B, Boyd R, Hamilton I, Oreszczyn T, et al. Health care's response to climate change: a carbon footprint assessment of the NHS in England. Lancet Planet Health. 2021;5(2):e84–92. https://doi.org/10.1016/S2542-5196(20)30271-0
89. Acemoglu D, Restrepo P. Automation and new tasks: how technology displaces and reinstates labor. J Econ Perspect. 2019;33(2):3–30. https://doi.org/10.1257/jep.33.2.3
90. World Economic Forum. The future of jobs report 2020. Geneva: World Economic Forum; 2020. https://www.weforum.org/publications/the-future-of-jobs-report-2020/
91. Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, et al. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. New York: ACM; 2020. p. 33–44. https://doi.org/10.1145/3351095.3372873
92. Shneiderman B. Human-centered artificial intelligence: reliable, safe & trustworthy. Int J Hum Comput Interact. 2020;36(6):495–504. https://doi.org/10.1080/10447318.2020.1741118

Figures

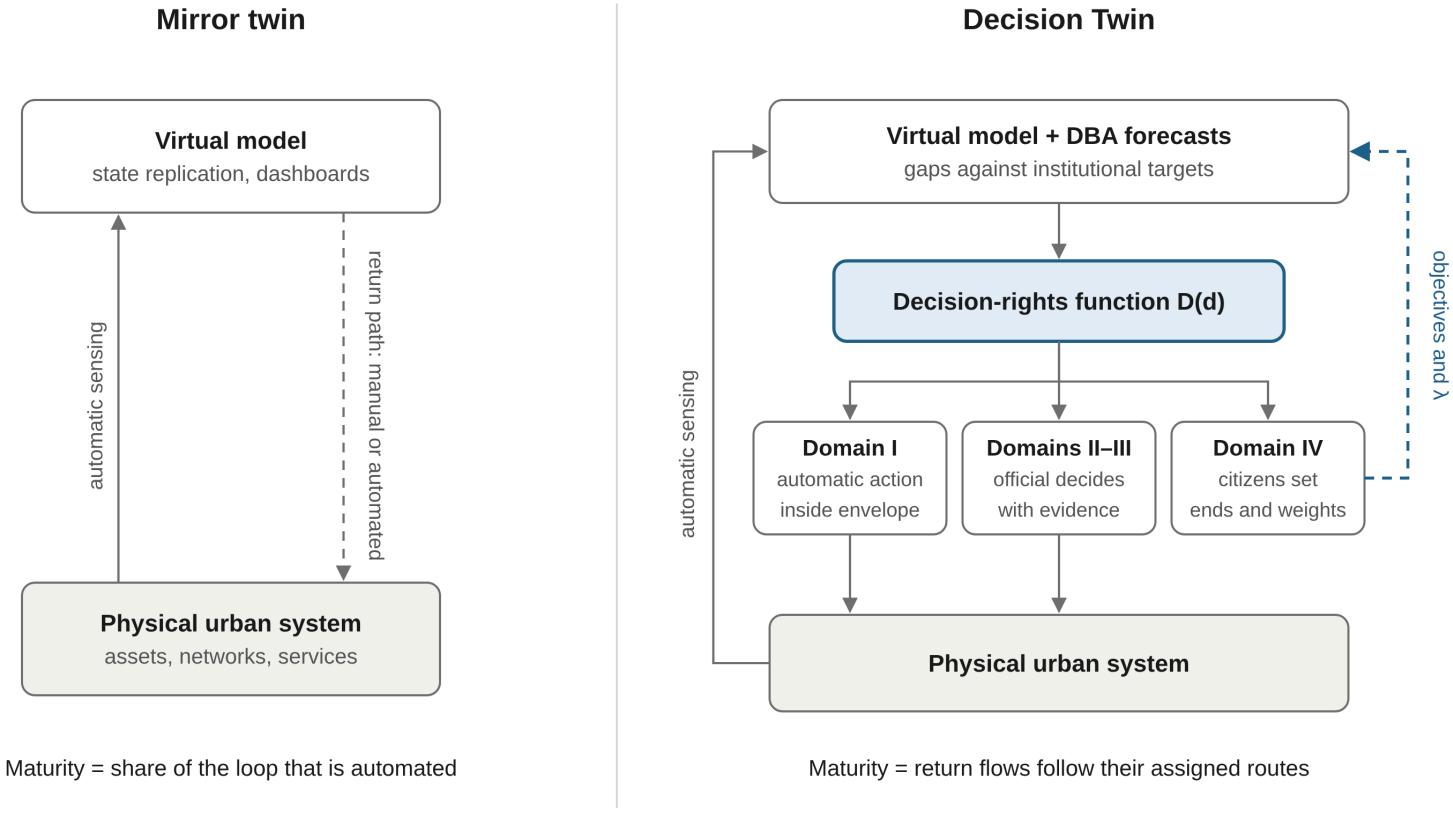

Figure 1

Mirror twin and Decision Twin compared. In a mirror twin the return path is a question of automation maturity; in a Decision Twin every return flow is routed by its assigned decision-rights domain

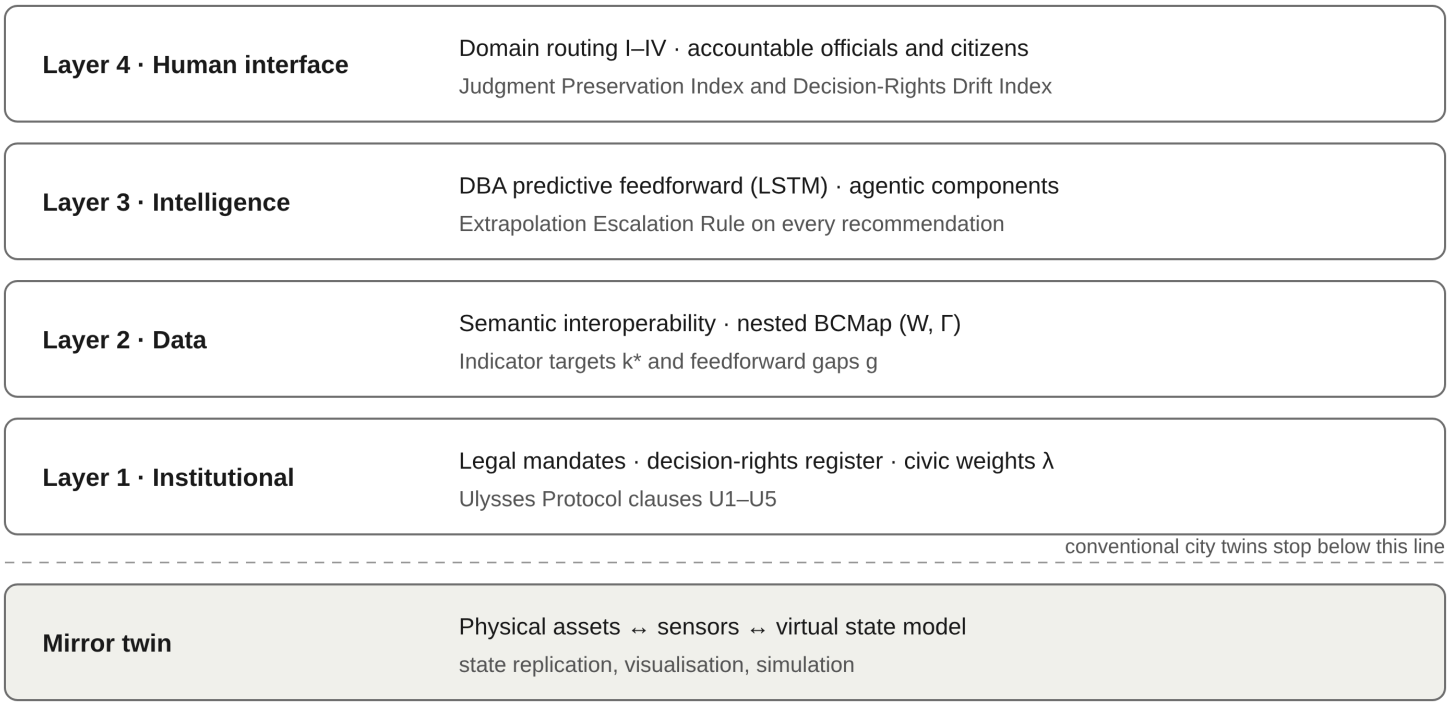

Figure 2

Decision Twin architecture: four governance layers of the Predictive Governance Architecture added to a conventional mirror twin

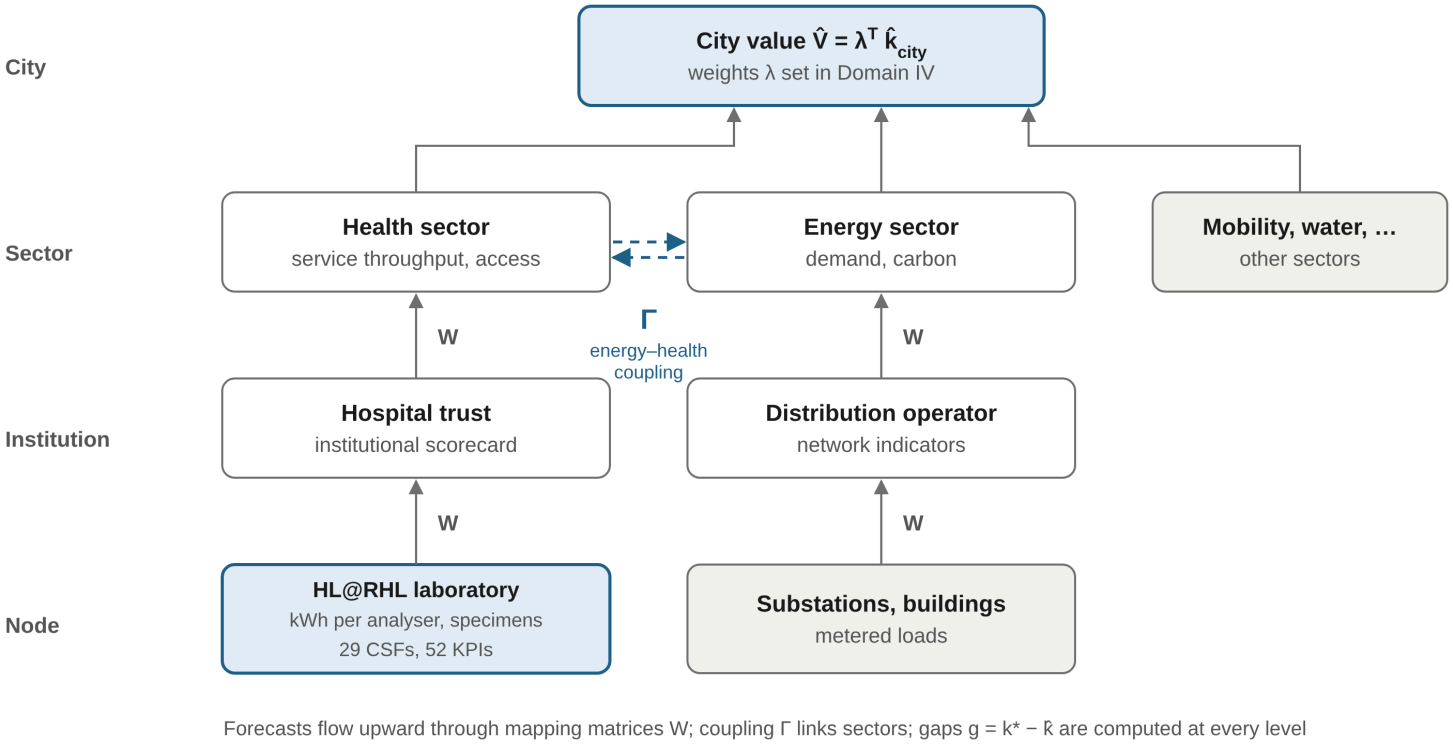

Figure 3

Nested BCMap: node forecasts composed into institutional, sector and city indicators through mapping matrices W, with cross-sector coupling Γ and civic weights λ

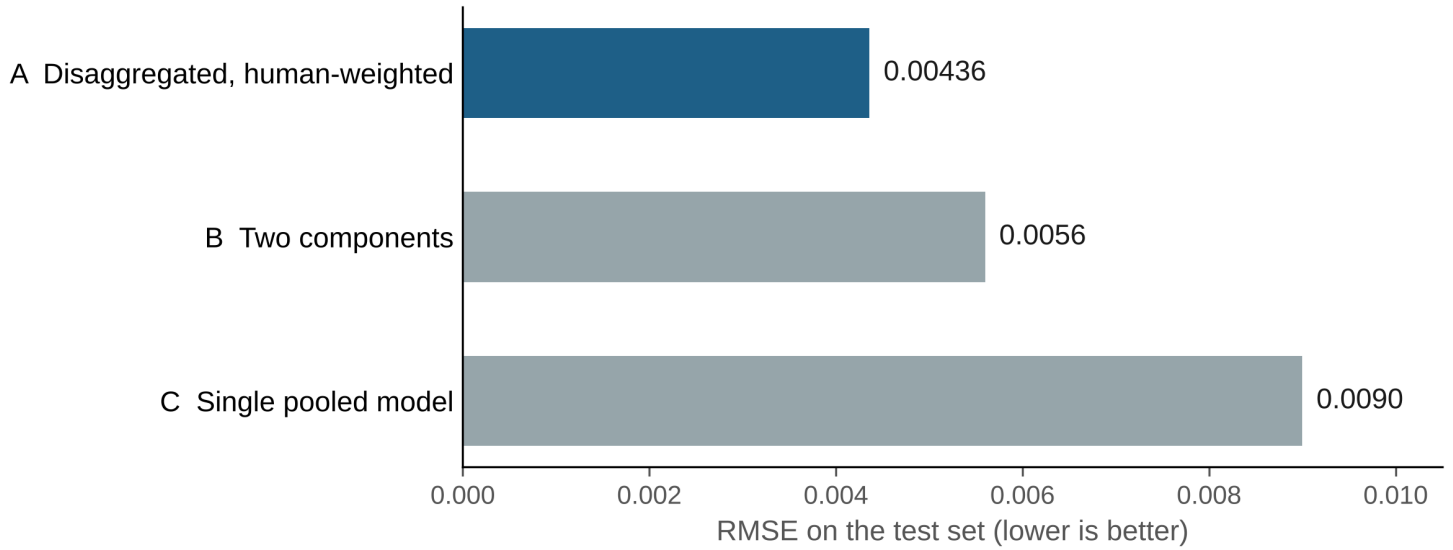

Figure 4

Forecast error of the three model configurations at HL@RHL (root mean squared error on the test set; lower is better). Data from [29]

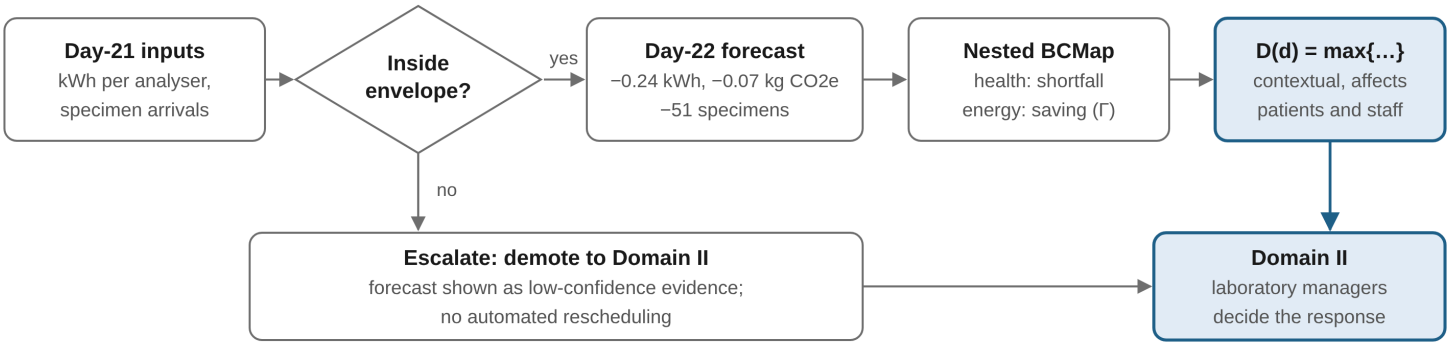

Figure 5

Illustrative routing of the HL@RHL day-ahead forecast through the Decision Twin, showing domain assignment and the Extrapolation Escalation Rule

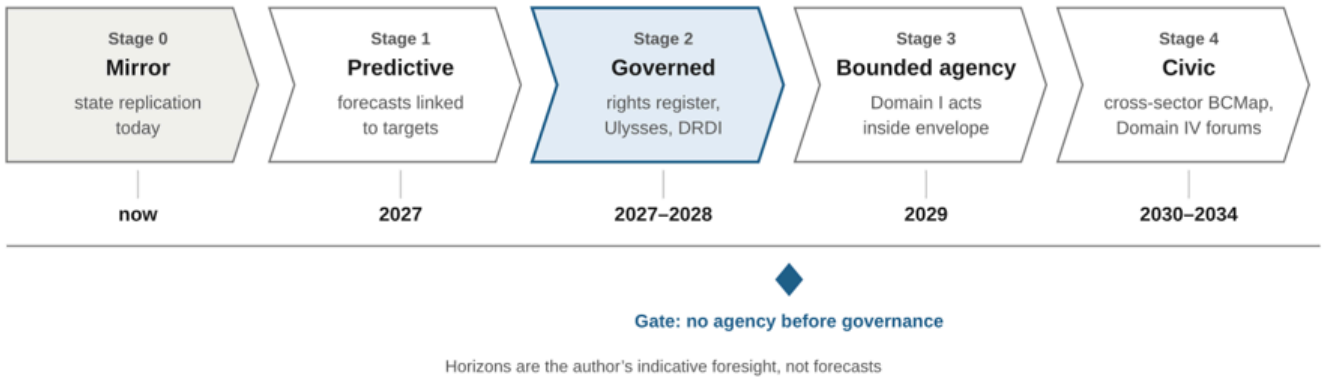

Figure 6

Decision Twin maturity pathway, with indicative foresight horizons for 2027–2034