

If artificial intelligence is used, who or what is really learning, contributing and solving problems? A case of indoor air



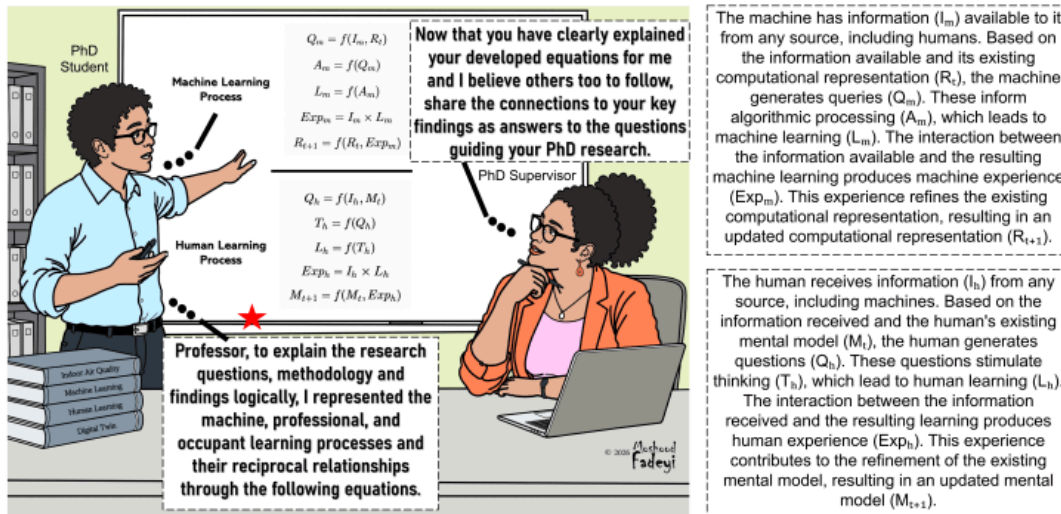
indooraircartoon.com/2026/10/26/if-artificial-intelligence-is-used-who-or-what-is-really-learning-contributing-and-solving-problems-a-case-of-indoor-air

26 October 2026

Indoor Air Cartoon Journal, October 2026, Volume 9, #183

[Cite as: Fadeyi MO (2026). If artificial intelligence is used, who or what is really learning, contributing and solving problems? A case of indoor air. *Indoor Air Cartoon Journal*, October 2026, Volume 9, #183.]

IF ARTIFICIAL INTELLIGENCE IS USED, WHO OR WHAT IS REALLY LEARNING, CONTRIBUTING AND SOLVING PROBLEMS? A CASE OF INDOOR AIR



PhD Student

Machine Learning Process

$$Q_m = f(I_m, R_t)$$

$$A_m = f(Q_m)$$

$$L_m = f(A_m)$$

$$Exp_m = I_m \times L_m$$

$$R_{t+1} = f(R_t, Exp_m)$$

Human Learning Process

$$Q_h = f(I_h, M_t)$$

$$T_h = f(Q_h)$$

$$L_h = f(T_h)$$

$$Exp_h = I_h \times L_h$$

$$M_{t+1} = f(M_t, Exp_h)$$

PhD Supervisor

Professor, to explain the research questions, methodology and findings logically, I represented the machine, professional, and occupant learning processes and their reciprocal relationships through the following equations.

Now that you have clearly explained your developed equations for me and I believe others too to follow, share the connections to your key findings as answers to the questions guiding your PhD research.

The machine has information (I_m) available to it from any source, including humans. Based on the information available and its existing computational representation (R_t), the machine generates queries (Q_m). These inform algorithmic processing (A_m), which leads to machine learning (L_m). The interaction between the information available and the resulting machine learning produces machine experience (Exp_m). This experience refines the existing computational representation, resulting in an updated computational representation (R_{t+1}).

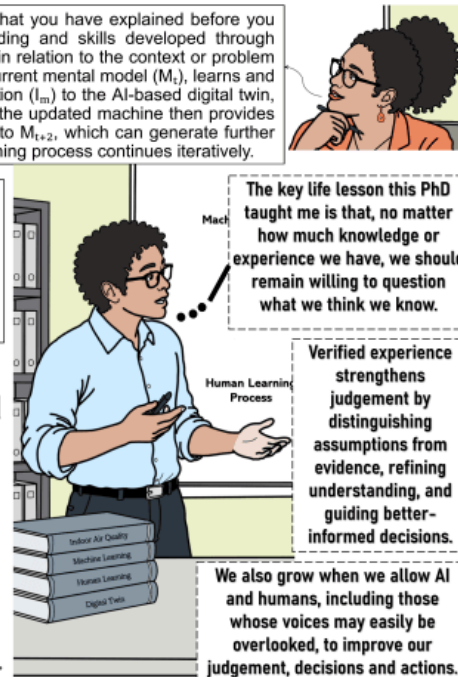
The human receives information (I_h) from any source, including machines. Based on the information received and the human's existing mental model (M_t), the human generates questions (Q_h). These questions stimulate thinking (T_h), which lead to human learning (L_h). The interaction between the information received and the resulting learning produces human experience (Exp_h). This experience contributes to the refinement of the existing mental model, resulting in an updated mental model (M_{t+1}).

What you've shared is quite original! Let me share my understanding of what you have explained before you move on to the results. Experience (Exp) is the knowledge, understanding and skills developed through learning, while a mental model (M) is the representation of that experience in relation to the context or problem at hand. A professional receives information (I_h), interprets it through their current mental model (M_t), learns and updates that mental model to M_{t+1} , and contributes new or revised information (I_m) to the AI-based digital twin, which learns and updates its computational representation from R_t to R_{t+1} ; the updated machine then provides new information (I_h) to the professional, supporting further human learning to M_{t+2} , which can generate further information for subsequent machine learning to R_{t+2} , and the reciprocal learning process continues iteratively.

An occupant receives information ($I_{h,o}$), interprets it through their current mental model ($M_{t,o}$), learns and updates that mental model to $M_{t+1,o}$, and contributes what they have learnt as new or revised information to the professional (I_h), the Artificial Intelligence-based digital twin (I_m), or both; this information can then support professional human learning from M_t to M_{t+1} , machine learning from R_t to R_{t+1} , or both pathways, and the resulting new information can return to the occupant as $I_{h,o}$, supporting further occupant learning to $M_{t+2,o}$, with the reciprocal learning process continuing iteratively.

Findings from RQ1 showed progressive AI-based digital-twin learning from independently verified indoor-air and other indoor-environmental problems, outperforming the frozen baseline; RQ2 findings showed deliberate professional learning and reciprocal professional–digital-twin learning improved performance beyond machine learning.

RQ3 findings showed occupant learning and participation moved occupants beyond complaints, contributing situated experience and value judgements that shaped decisions and actions in the reciprocal occupant–professional–AI-based digital twin problem-solving system.



PhD Student

Machine Learning Process

$$Q_m = f(I_m, R_t)$$

$$A_m = f(Q_m)$$

$$L_m = f(A_m)$$

$$Exp_m = I_m \times L_m$$

$$R_{t+1} = f(R_t, Exp_m)$$

Human Learning Process

$$Q_h = f(I_h, M_t)$$

$$T_h = f(Q_h)$$

$$L_h = f(T_h)$$

$$Exp_h = I_h \times L_h$$

$$M_{t+1} = f(M_t, Exp_h)$$

PhD Supervisor

Professor, to explain the research questions, methodology and findings logically, I represented the machine, professional, and occupant learning processes and their reciprocal relationships through the following equations.

Now that you have clearly explained your developed equations for me and I believe others too to follow, share the connections to your key findings as answers to the questions guiding your PhD research.

The machine has information (I_m) available to it from any source, including humans. Based on the information available and its existing computational representation (R_t), the machine generates queries (Q_m). These inform algorithmic processing (A_m), which leads to machine learning (L_m). The interaction between the information available and the resulting machine learning produces machine experience (Exp_m). This experience refines the existing computational representation, resulting in an updated computational representation (R_{t+1}).

The human receives information (I_h) from any source, including machines. Based on the information received and the human's existing mental model (M_t), the human generates questions (Q_h). These questions stimulate thinking (T_h), which lead to human learning (L_h). The interaction between the information received and the resulting learning produces human experience (Exp_h). This experience contributes to the refinement of the existing mental model, resulting in an updated mental model (M_{t+1}).

The key life lesson this PhD taught me is that, no matter how much knowledge or experience we have, we should remain willing to question what we think we know.

Verified experience strengthens judgement by distinguishing assumptions from evidence, refining understanding, and guiding better-informed decisions.

We also grow when we allow AI and humans, including those whose voices may easily be overlooked, to improve our judgement, decisions and actions.

REFERENCES

- Hemmer, P., Schemmer, M., et. al. (2025). *European Journal of Information Systems*, 34(6), 979-1002.
 Mahajan, S., Mondardini, R., and Helbing, D. (2024). *Sustainable Cities and Society*, 116, 105890.
 Xie, X., Merino, J., Moretti, N., et. al. (2023). *Automation in Construction*, 146, 104695.

© Moshood O. Fadeyi 2026 | Contact at: mofadeyi@gmail.com | www.indooraircartoon.com

Fictional Case Story (Audio – available online)– Part 1 (Preface, Ch 1 & Ch 2)

Fictional Case Story (Audio – available online)– Part 2 (Ch 3)

Fictional Case Story (Audio – available online)– Part 3 (Ch 3 cont'd)

Fictional Case Story (Audio – available online) – Part 4 (Ch 3 cont'd)

Fictional Case Story (Audio – available online) – Part 5 (Ch 4)

Fictional Case Story (Audio – available online) – Part 6 (Ch 4 cont'd)

Fictional Case Story (Audio – available online) – Part 7 (Ch 4 cont'd)

Fictional Case Story (Audio – available online) – Part 8 (Ch 5 & Ch 6)

..... Preface

Across education, research, professional practice, industry and everyday life, mechanistic externalisation through examinations, writing, calculations, presentations and other visible outputs had long been treated as a value-adding activity, rather than recognised as necessary non-value-added work whose human effort could be streamlined through technological advancement. As humans were expected to perform this mechanistic externalisation themselves, the quality of the visible output became closely associated with their capability and value. Less visible cognitive activities, including learning, questioning, critical and reflective thinking, logical deduction, abstract reasoning and creative imagination, which inform judgement and decision-making and guide what needs to be externalised, therefore received less recognition despite governing value-oriented problem diagnosis and solving. For people who struggled with mechanistic externalisation, others could wrongly interpret that difficulty as evidence of weaker intellectual and problem-solving capability, and over time the individuals themselves could begin to believe the same thing.

Artificial intelligence unsettled this relationship. As a result of established practices and perceptions before AI, when AI digitally externalised an output, people could assume that the human had contributed little to producing it. On the other hand, people could delegate externalisation and much of their cognitive governance to AI while still producing polished outputs that were not necessarily effective. For one building engineer who had long struggled with mechanistic externalisation, this problem became personal and led him to investigate human-machine learning through value-oriented indoor-air problem diagnosis. His journey forms the subject of this fiction story.

..... Chapter 1

Society and education systems often mistake performance in mechanistic externalisation, although necessary but non-value-added work, for the true measure of problem-solving capability, while giving insufficient attention to developing the value-adding cognitive capabilities to learn, think, form informed judgements and make decisions that guide value-oriented problem diagnosis and solving. Mustafa Abdul Matin did not understand this distinction when he was growing up. What he understood was much simpler and more painful: some children appeared naturally intelligent, while he did not seem to be one of them. Over time, that appearance became part of how other people saw him and, eventually, how he saw himself. Mustafa grew up in a family where education mattered greatly. His father, Abdul Rahman, believed education could open doors that hardship had kept closed for earlier generations, while his mother, Mariam, believed persistent effort would eventually lead somewhere. His older sister, Safiya, performed strongly in school,

while his younger brother, Yusuf, could explain what he had learnt with confidence. Nobody deliberately created a competition among them, but marks, report cards, prizes and teachers' comments became the visible evidence through which ability was discussed. Mustafa slowly learnt to use the same evidence to judge himself.

During primary school, his marks usually sat around the middle or lower half of the class, often in the 50 to 60 per cent range. Written subjects could fall into the forties when he understood an idea but could not express it in the expected form. Science and mathematics were sometimes better, particularly when questions allowed him to reason from relationships, but he was inconsistent because one difficult question could hold his attention for too long. While other pupils answered quickly, Mustafa often found that one question generated several more in his mind. If a teacher asked why a plant beside a window grew towards the light, Mustafa might first think of sunlight, then wonder what would happen if there were several windows, if the light moved, or if mirrors redirected it. Before he could organise these thoughts, another pupil would answer, "Plants need sunlight," receive praise, and the class would move on. Mustafa remained silent, not because he had nothing to contribute, but because what was happening inside his mind had not become words quickly enough. He did not yet understand that the speed of externalisation and the quality of the thinking being externalised were not the same thing.

At home, Safiya sometimes returned with marks in the seventies or eighties and enthusiastic comments from her teachers. Abdul Rahman would congratulate her and tell the younger children to learn from their sister. When Mustafa brought home another paper filled with marks in the fifties, his father would ask, "You knew some of this. Why did you not write it properly?" Mustafa had no language for the gap between understanding and expression. He gradually concluded that if he could not explain something clearly, perhaps he did not really understand it. Other people's judgement slowly became his own. By lower secondary school, that self-image had hardened. Written answers became longer, examinations more consequential and the pressure to retrieve, organise and express information quickly more intense. Mustafa passed most subjects but rarely achieved distinctions, and language-heavy subjects remained particularly difficult. Yet one science lesson offered evidence that did not fit the story he was telling himself. The teacher described a bedroom that remained warmer than other rooms even though all were served by the same air-conditioning system. Several pupils suggested a faulty air conditioner or a thermostat adjustment.

When the teacher asked Mustafa, he wanted to know whether the room faced the afternoon sun, whether its windows were larger, whether occupancy was different, whether the door stayed closed and whether the room actually received the same quantity of cool air. Perhaps, he suggested, the problem was not simply the air conditioner. The room might be gaining more heat, receiving less cooling, or both. The teacher praised his reasoning, but a week later another ordinary examination result arrived. Mustafa treated the classroom episode as an exception and the grades as truth. Safiya's educational

journey made the comparison sharper. After four years of secondary school, she entered a two-year A-level programme and then university to study Business Administration. Her route looked like the route designed for academically strong students. Mustafa's final secondary-school results allowed him to continue his education, but not through the same direct path. While many classmates entered two years of A-level study and others entered three years of polytechnic education, Mustafa was placed on a three-year post-secondary technical route.

He knew the technical route was legitimate, but emotionally it felt like an official ranking of capability. Yet the environment also exposed him to another kind of learning. Knowledge was connected to systems, equipment, procedures and things that had to work in practice. He still struggled with examinations and written work, but practical problems allowed him to ask questions that did not emerge as naturally in conventional assessment. Without recognising it, he was learning to think about how parts of a system interacted. After those three years, Mustafa progressed to a polytechnic and began another three-year diploma. By then, classmates who had taken A-levels were already at university, while those who had entered polytechnic directly were completing their diplomas. Mustafa had spent three years simply reaching the point at which some of them had started. His diploma suited some aspects of his thinking better because theory was tied to practical applications, yet reports, presentations and timed examinations remained difficult. His results were respectable enough for progression, but not enough to change how he saw himself.

Yusuf, although younger, followed a shorter route. After secondary school he entered a three-year polytechnic programme in Information Technology and later progressed to university to study Computing. He never belittled Mustafa and continued seeking his advice on difficult personal or technical decisions, but Mustafa found the comparison painful. Within the same family, Safiya had moved through A-levels to university, Yusuf through polytechnic to university, while Mustafa had needed three years of technical education followed by three years of polytechnic before he could even seek university admission seriously. Family gatherings amplified the difference. Relatives would ask about former classmates who were already at university and then turn to Mustafa with the innocent question, "So, what are you doing now?" Abdul Rahman would say everyone had a different journey, while Mariam reminded others that Mustafa never stopped trying. Mustafa appreciated their support, but persistence felt like a consolation when what he wanted was evidence that he was capable. Six post-secondary years had passed before he could apply to university, compared with two years for many A-level classmates and three for those who entered polytechnic directly.

Mustafa applied to study Architecture. He imagined it could combine imagination, technical understanding and his persistent interest in relationships between people, buildings and problems. His application was rejected. After six years of post-secondary education, the rejection felt larger than one institutional decision; it seemed to confirm a judgement he had carried since childhood. Abdul Rahman asked whether there was an appeal process.

Mariam reminded him that one rejection should not erase six years of effort. Safiya helped him review the appeal requirements, and Yusuf told him that after travelling that far he had nothing to lose by trying once more. The appeal succeeded, but the university did not admit him to Architecture. It offered Building Engineering instead. Mustafa privately saw it as a consolation: the university had still not considered him suitable for what he really wanted, but had found somewhere else to place him. Abdul Rahman told him, “You asked for a way into the university, and a way has opened. Do not decide what this course can become before you have entered it.” Mustafa accepted, although he felt he had entered through a side door rather than the front entrance used by stronger students.

University did not remove his old difficulty. Reading loads increased, assignments required more sophisticated written arguments, examinations demanded faster organisation of complex information, and presentations exposed the gap between what he thought and what he could communicate. He could spend twenty minutes on one paragraph because he wanted to know what the author meant, what assumptions supported the argument and under what conditions the conclusion might fail. That habit deepened his understanding but slowed his visible output. He often saw the unfinished document rather than the thinking that had produced the delay. When he graduated, Mustafa received a Second Class Lower degree. He accepted that the classification was the legitimate outcome of the university’s assessment system, but he also felt that it did not align with the depth of ideas, questions and connections in his mind. His family saw the degree before the classification. Abdul Rahman and Mariam were proud that the son who had travelled through technical education, polytechnic, rejection and appeal had graduated. Safiya congratulated him without qualification, while Yusuf joked that the brother everyone approached with difficult problems finally had the certificate he had chased for years. Mustafa smiled, yet “Second Class Lower” felt like an official ceiling.

Years later, Mustafa would recognise the irony of Building Engineering. The programme he had regarded as a consolation had placed him inside the study of buildings as interacting systems. He learnt about heat transfer, ventilation, airflow, building services and environmental performance. Combined with technical and polytechnic experience, it strengthened his tendency to ask how different conditions interacted rather than accepting the most visible variable as the complete explanation. Architecture had been the door he wanted, but Building Engineering became part of the route through which he learnt to investigate why buildings performed as they did and how that performance affected people. Artificial intelligence was not present during Mustafa’s primary, secondary, technical, polytechnic or university education. Every examination, report and presentation through which he had been judged depended on his own ability to externalise what was inside his mind. After graduation, however, the same difficulty followed him into industry. Workplaces used different language, but recognition still depended heavily on reports, presentations, meetings, performance reviews and the ability to make one’s contribution visible to managers and decision-makers.

Mustafa could manage difficult work. When something went wrong, he wanted to know what had changed, what information was missing, whether the obvious explanation was really the cause and what consequences a proposed solution might create elsewhere. Colleagues increasingly approached him when a problem refused to fit its first explanation. Yet he remained poor at turning such contributions into polished narratives of achievement. Others could present confidently, summarise accomplishments neatly and position themselves visibly for promotion. Mustafa's promotions came slowly, and sometimes not at all. One office investigation stayed with him. Occupants complained of headaches and discomfort while measurements showed elevated carbon dioxide. The immediate recommendation was to increase outdoor-air ventilation. Mustafa asked when complaints occurred, whether everyone was affected, whether they were concentrated in one area, whether carbon dioxide followed the same pattern and what had changed before the complaints began. Renovation had recently taken place nearby, so he asked whether emissions from new materials had been investigated. The eventual explanation involved occupancy, ventilation distribution and material emissions rather than one convenient cause.

Mustafa had reframed the problem, separated observation from cause, identified missing information and generated competing explanations. Yet he still regarded this as merely doing his job. The absence of promotion reinforced the same message he had absorbed from education: capability seemed to belong to people who could express, present and promote their work convincingly. Industry appeared to value many of the same visible forms of performance that education had rewarded. The years of struggle, however, were developing something Mustafa had not named. Because answers rarely came easily, he had learnt to question rather than accept. Because externalisation was difficult, he often spent longer checking whether an idea was logically coherent. Because he had repeatedly suffered from misunderstanding questions or failing to communicate what he meant, he became attentive to context, ambiguity, missing information and the difference between what was known, what was assumed and what still needed investigation. Without realising it, he was developing cognitive governance: the capability to regulate his thinking, question his mental models, examine the basis of a judgement, recognise uncertainty and revise his understanding.

His long route also forced him to become a lifelong learner. When one educational route closed, he learnt through another. When existing knowledge did not explain a professional problem, he searched for more. When experience contradicted what he thought he knew, he reconstructed his understanding. What Mustafa had experienced mainly as struggle had also been training him to remain cognitively functional when familiar answers were insufficient. Many years after graduation, artificial intelligence began entering everyday intellectual and professional work. Mustafa quickly discovered that it could search, summarise, calculate, organise information, draft text and produce polished language with remarkable speed. For generations, exceptional skill in such tasks had given some people

an important advantage because there had been no practical alternative to doing them personally. AI began disrupting that advantage by automating or supplementing many forms of externalisation.

For Mustafa, the change was profound. He could describe an idea imperfectly, examine what the machine produced and decide whether the externalised form represented his understanding. Yet he also discovered that AI could produce beautifully written nonsense. It could frame the wrong problem, introduce unsupported causal relationships, overlook stakeholders or recommend technically plausible actions that failed in context. Mustafa therefore did what he had been doing for years: he questioned. He challenged assumptions, introduced missing context, rejected answers that confused symptoms with causes and repeatedly revised the interaction until the output became more defensible. The very habits that once made him appear slow now gave him an advantage. He was less impressed by a response merely because it sounded intelligent. AI could supply speed, alternatives, calculations and language, while Mustafa's value increasingly lay in governing the interaction: deciding what should be asked, recognising when the question itself was wrong, judging whether evidence was credible and refusing to stop simply because an answer looked complete.

One evening Safiya watched him reject and revise several polished AI-generated paragraphs. "If it can write so well, why do you keep changing what it writes?" she asked. Mustafa replied, "Because writing it well does not mean it has understood what I am trying to solve." The sentence captured a distinction he had spent much of his life failing to see. Externalising thinking and determining the quality of the thinking being externalised were related, but they were not identical. AI had altered the value of different capabilities. Externalisation could increasingly be supplemented by machines, while the scarce capability shifted towards governing thinking and learning. This did not mean that people who had previously excelled academically lacked deeper capability, nor that Mustafa had suddenly become superior. It meant that those who had developed strong externalisation without equally developing questioning, testing and reflective judgement could now find that AI replicated much of what had distinguished them. Mustafa could use AI to supplement his greatest weakness while bringing to the interaction cognitive governance and lifelong learning capability that had developed over decades.

He also noticed a social contradiction. Before AI, people had often treated polished externalised work as evidence of the human's intellectual capability. Once AI could assist with that externalisation, some people quickly concluded that "AI did the work." The human contribution was judged by how much writing, searching, calculating or drafting the person had personally performed rather than by the capability demonstrated in learning, thinking, questioning, interpreting evidence, forming informed judgements, making decisions and guiding value-oriented problem diagnosis and solving. Mustafa did not deny the risks. People could use AI to avoid learning, accept incorrect outputs or disguise ignorance behind polished language. But the presence of AI alone could not settle the attribution

question. Two people could use the same system and obtain very different value. One might accept the first fluent answer; another might recognise faulty framing, challenge assumptions, provide missing evidence, test alternatives and continue learning from the interaction. Access to the same AI did not equalise the quality of judgement.

In that sense, AI reversed part of Mustafa's lifelong disadvantage. The capability in which he had been weakest, externalisation, could now be substantially supplemented.

Meanwhile, the questioning, thoroughness, logical deduction, reflective thinking, cognitive governance and lifelong learning capability he had unintentionally developed became increasingly valuable. What once looked inefficient could now protect him from being impressed by speed, fluency and apparent certainty. AI had not made him capable; it had made one of his limitations less dominant and revealed the value of capabilities that had long been hidden behind it.

..... Chapter 2

Several years after his first degree, Mustafa returned to university for a part-time Master of Science while continuing to work full time. The experience was radically different from his earlier education. Assessment relied more heavily on coursework, applied analysis, projects and sustained assignments that allowed students time to investigate, connect ideas and revise reasoning. Timed examinations no longer dominated his experience of what counted as learning. The same thoroughness that once left examination questions unfinished could now improve a final submission. Professional experience also changed how he learnt. Concepts were no longer isolated material to remember; he could connect them with buildings, systems, occupants and real problems. AI was now available to support externalisation, but Mustafa did not treat its first response as a destination. He checked outputs against sources and professional knowledge, questioned weaknesses and used disagreement as a reason to investigate further. The MSc created an unusual alignment between his learning style and the assessment system: he had room to question, reflect, integrate and revise, while AI reduced some of the friction involved in making that thinking visible.

The results were unlike anything in his earlier academic life. Module after module, strong grades accumulated. Mustafa eventually graduated with distinction and a Grade Point Average of 4.93 out of 5.00. The same person who had needed three years of technical education and three years of polytechnic before barely entering university, who had gained admission only after appeal and graduated with Second Class Lower, had now achieved 4.93 out of 5 in a demanding MSc. Mustafa did not use the result to declare that his earlier education had been wrong or that he had suddenly become intellectually superior. He had grown through professional practice, reading and experience, and that development mattered. Yet the contrast also showed that assessment conditions mattered. The MSc gave greater space to sustained questioning, reflective thinking, logical deduction,

integration of knowledge and iterative refinement of judgement. Different assessment systems could reveal different dimensions of a developing person's capability.

Abdul Rahman struggled to hide his satisfaction when Mustafa shared the final GPA. Mariam repeated, "4.93 out of 5?" as though she needed confirmation. Safiya reminded him how many years he had used stronger-performing people as evidence against himself, while Yusuf asked whether he was finally ready to admit that examinations had never told the entire story. Mustafa smiled, but remained cautious. If Second Class Lower had not defined the whole of him, neither could 4.93. After the MSc, Mustafa was offered a scholarship to pursue a PhD full time. By then he was married. His wife, Amina, was a secondary-school teacher who understood both the opportunity and the financial risk. The scholarship helped, but leaving full-time employment meant the household would sacrifice income. Amina agreed to carry much of the financial burden through her salary, giving Mustafa the practical freedom to undertake the PhD. He never regarded her contribution as incidental; even work carrying one person's name could depend on sacrifices and support invisible in the final document.

As he shaped his doctoral direction, attitudes towards AI in education and professional practice increasingly troubled him. Some people embraced AI uncritically, while others treated its use almost automatically as evidence that the human had not done the real work. Mustafa heard academics question whether a student's work could genuinely be theirs if AI had helped structure or express it, and professionals dismiss strong outputs by saying, "AI did it." Such reactions returned him to the attribution problem that had followed him throughout his life. He became less interested in defending either humans or machines and more interested in tracing contribution. If AI produced an answer but a professional recognised that the problem had been framed incorrectly, who had contributed what? If professional context changed the machine's analysis, where had learning occurred? If the machine revealed information the professional had overlooked and the professional revised a judgement, had AI replaced human learning or stimulated it? If both changed through interaction, calling the result simply "AI work" or "human work" seemed intellectually inadequate.

Mustafa needed a real domain in which such questions mattered. Indoor air quality attracted him because the interest had roots in Building Engineering and professional practice. Buildings had taught him about ventilation, airflow, thermal conditions and environmental performance, while practice had shown him that occupants experienced buildings through stuffiness, odours, draughts, heat, noise, discomfort and uncertainty about whether the environment around them was healthy or safe. Indoor-air problems also resisted superficial diagnosis. A complaint could be real while an occupant's explanation of its cause was wrong. A sensor could measure something accurately without that measurement being sufficient to explain the problem. A building-management system could show equipment operating normally while occupants remained uncomfortable. A technically effective intervention could solve one problem while creating another

elsewhere. Hospitals made these tensions even more consequential because ventilation, filtration, pressure relationships, temperature control, clinical activities, patients, staff and visitors continuously interacted.

Mustafa saw in indoor-air and other indoor-environmental problems the same intellectual challenge that had followed him since childhood. The first visible symptom might not be the problem. The first plausible explanation might not be the cause. The most confident person might not possess the most complete understanding, and the most polished machine output might still be wrong. Useful diagnosis required questioning, integration, verification, judgement and learning from what happened afterwards. He began reading about digital twins, artificial intelligence, building monitoring, fault diagnosis, professional decision-making, human learning and occupant participation. Machines, professionals and occupants each possessed access to different information, and each could also be wrong. A machine could learn from poor information, a professional could be biased by previous experience, and an occupant could correctly perceive a problem while incorrectly interpreting its cause. The important issue was therefore not merely collecting more information, but understanding how information could be questioned, verified, learnt from and carried forward.

The pieces of Mustafa's journey finally converged. The child who struggled to externalise thought, the technical student who took the longer route, the Building Engineering graduate with Second Class Lower, the professional who struggled for promotion, the AI user whose old weakness could now be supplemented, the part-time MSc student who graduated with 4.93 out of 5, the husband supported by Amina while pursuing a full-time PhD, and the researcher fascinated by indoor air were no longer separate versions of the same man. They had become parts of one inquiry. His doctoral research would not attempt to prove that humans were superior to artificial intelligence, nor that AI could replace human judgement. Mustafa wanted something more demanding: to investigate what machines, professionals and eventually occupants could actually learn and contribute, how their capabilities could supplement one another, where each could fail, and how their combined learning might support better value-oriented problem diagnosis and solving. Only then was he ready to formulate the practical problem his PhD would investigate.

"Large hospitals depend on complex ventilation, air-conditioning, pressure-control and other building systems to maintain indoor environments that support health, comfort, clinical work and safe operation. In practice, however, indoor-air and other indoor-environmental problems may still recur. Occupants may report odour, stuffiness, thermal discomfort, draught, noise or other undesirable conditions, while environmental monitoring or building-management systems may indicate abnormal pollutant concentrations, pressure relationships or equipment operation. The practical challenge is not merely recognising that something is wrong. It is correctly diagnosing why the problem is occurring and determining what should be done to provide value to the affected stakeholders without creating unacceptable consequences elsewhere in the hospital.

The current performance situation is that diagnosis and solving can be fragmented, reactive and dependent on the information and experience available at the time of each problem. Environmental measurements, building-management-system data, maintenance records, spatial relationships, professional knowledge and occupant experience may exist, but they are not necessarily brought together or learnt from systematically. A recurring problem may therefore be approached largely as a new problem even when related problems have previously been investigated. Possible causes may be confused with symptoms, important competing explanations may be overlooked, further information may not be requested when needed, and technically plausible solutions may fail to account adequately for clinical workflow, comfort, convenience, safety, operational constraints or stakeholder sacrifices.

Digital twins and AI offer the possibility of analysing large amounts of building and environmental information, but a digital twin that does not progressively learn from independently verified problem outcomes remains limited by its existing computational representation. Conversely, machine learning alone may also be insufficient. Professionals possess contextual knowledge, practical judgement and experience that may not be represented computationally, but professional learning may remain individual, informal or disconnected from subsequent machine learning. Occupants possess situated experience of where, when and under what circumstances environmental problems are experienced, yet their role is often limited to reporting complaints rather than participating meaningfully in diagnosis and solving. Their observations may therefore be underused, while their interpretations may also be accepted too readily without independent verification.

The targeted performance situation is an indoor-environmental problem-solving system that becomes progressively more capable through verified experience. Such a system should be able to identify and compare plausible causes, recognise uncertainty, determine what additional information is required, distinguish symptoms from underlying causes, and assess solutions in terms of both technical effectiveness and stakeholder value. It should also enable professionals to develop their own questioning, reasoning and judgement rather than become dependent on AI outputs, and enable occupants to contribute relevant situated experience without treating subjective interpretation as technical truth. In this targeted situation, machine learning, professional human learning and occupant learning would contribute to one another through verified information while remaining subject to professional governance and safety constraints.

The practical problem is therefore the gap between the current fragmented and partly static problem-solving capability and the targeted capability of an integrated, progressively learning, value-oriented system. Several potential barriers may be responsible for this gap. These include incomplete or poorly integrated information; inability to retain and reuse verified experience; weak distinction between observations, symptoms, hypotheses and causes; limited transfer of professional learning into machine learning; limited transfer of machine-generated information into professional learning; underuse of occupant situated

experience; conflicting or unreliable human interpretations; multi-causal, intermittent and spatially complex problems; time pressure; uncertainty; and the possibility that previous experience may introduce confirmation, automation or experience-related bias.

An effective solution therefore cannot be designed simply by adding more sensors, introducing AI, involving professionals or collecting more occupant feedback. The root causes and boundaries of the performance gap must first be established empirically. It is necessary to determine whether an AI-based digital twin can learn progressively from independently verified problems, whether deliberate professional learning and reciprocal professional human–AI-DT interaction can improve capability beyond machine learning alone, and whether developing occupant learning, participation and agency can provide further value beyond the professional human–AI-DT system. The research was therefore required to determine what should learn, from whom, through what information pathways, under what conditions, and with what limits, so that a solution capable of moving hospital indoor-environmental problem diagnosis and solving from the current performance situation towards the targeted performance situation could be developed.”

This need forms the basis for the research questions and hypotheses that guided his PhD study.

(i) To what extent can an AI-based digital twin progressively learn to frame, diagnose and support the solving of recurring indoor-air and other indoor-environmental-condition problems in a large general hospital? Under what problem, environmental, spatial, temporal and operational conditions can it support value-oriented problem diagnosis and solving when the learning of hospital indoor-environmental problem-solving professionals is not deliberately developed?

(ii) How does deliberate learning by hospital indoor-environmental problem-solving professionals influence value-oriented diagnosis and solving of recurring indoor-air and other indoor-environmental-condition problems when they interact with a learning AI-based digital twin? To what extent, through what mechanisms, and under what conditions does reciprocal human–AI-DT learning improve value-oriented problem diagnosis and solving beyond AI-DT learning without deliberately developed professional human learning?

(iii) How can a reciprocally learning human–AI-DT system enable and empower hospital occupants to participate meaningfully in value-oriented diagnosis and solving of indoor-air and other indoor-environmental-condition problems? To what extent, through what mechanisms, and under what conditions does incorporating occupants’ situated experiences, questions, judgements and learning improve problem diagnosis and solving beyond the professional human–AI-DT learning system alone?

For the first research question, the Null Hypothesis (H_{01}) is that progressive AI-DT learning, without deliberately developing corresponding learning among hospital indoor-

environmental problem-solving professionals, does not produce meaningful improvement in prespecified dimensions of value-oriented diagnosis and solving of recurring indoor-air and other indoor-environmental-condition problems. Its performance does not meaningfully improve across the investigated problem, environmental, spatial, temporal and operational conditions relative to a non-learning digital-twin baseline. The Alternative Hypothesis (H_{11}) is that progressive AI-DT learning, without deliberately developing corresponding learning among hospital indoor-environmental problem-solving professionals, produces meaningful improvement in one or more prespecified dimensions of value-oriented diagnosis and solving of recurring indoor-air and other indoor-environmental-condition problems. The magnitude of improvement varies systematically with identifiable problem, environmental, spatial, temporal or operational conditions.

For the second research question, the Null Hypothesis (H_{02}) is that deliberately coupling professional human learning with AI-DT learning does not produce meaningful improvement in prespecified dimensions of value-oriented diagnosis and solving of recurring hospital indoor-air and other indoor-environmental-condition problems beyond AI-DT learning without deliberately developed professional human learning. No reproducible additional benefit is attributable to reciprocal development of the professional human mental model and AI-DT computational representation under the investigated conditions. The Alternative Hypothesis (H_{12}) is that under identifiable problem, environmental, spatial, temporal, operational or decision conditions, deliberately coupling professional human learning with AI-DT learning produces meaningful improvement in one or more prespecified dimensions of value-oriented diagnosis and solving of recurring hospital indoor-air and other indoor-environmental-condition problems beyond AI-DT learning without deliberately developed professional human learning. The improvement is associated with identifiable reciprocal learning mechanisms between development of the human mental model and development of the AI-DT computational representation.

For the third research question, the Null Hypothesis (H_{03}) is that enabling occupants to participate through their situated experiences, questioning, judgement and learning does not produce meaningful improvement in prespecified dimensions of value-oriented diagnosis and solving of recurring hospital indoor-air and other indoor-environmental-condition problems beyond the professional human–AI-DT learning system alone. It also does not produce meaningful improvement in prespecified dimensions of occupant capability or agency to participate in such diagnosis and solving. The Alternative Hypothesis (H_{13}) is that under identifiable problem, occupant, environmental, spatial, temporal and operational conditions, enabling occupants to participate through their situated experiences, questioning, judgement and learning produces meaningful improvement in one or more prespecified dimensions of value-oriented diagnosis and solving beyond the professional human–AI-DT learning system alone. It also strengthens one or more prespecified dimensions of occupants' capability or agency to participate meaningfully in indoor-environmental problem diagnosis and solving.

The research problem and questions informed the following objectives of his PhD study:

- (i) To examine the extent to which an AI-based digital twin can progressively learn to frame, diagnose and support the solving of recurring indoor-air and other indoor-environmental-condition problems in a large general hospital. This includes identifying the problem, environmental, spatial, temporal and operational conditions under which it can support value-oriented problem diagnosis and solving when professional human learning is not deliberately developed.
- (ii) To examine how deliberate learning by hospital indoor-environmental problem-solving professionals influences value-oriented diagnosis and solving of recurring indoor-air and other indoor-environmental-condition problems when interacting with a learning AI-based digital twin. This includes determining the extent, mechanisms and conditions under which reciprocal professional human–AI-DT learning improves value-oriented problem diagnosis and solving beyond AI-DT learning without deliberately developed professional human learning.
- (iii) To examine how a reciprocally learning professional human–AI-DT system can enable and empower hospital occupants to participate meaningfully in value-oriented diagnosis and solving of indoor-air and other indoor-environmental-condition problems. This includes determining the extent, mechanisms and conditions under which occupants’ situated experiences, questions, judgements and learning improve problem diagnosis and solving beyond the professional human–AI-DT learning system alone.

..... Chapter 3

Research Methods

Methods for Research Question 1:

Overview

The study was designed as a prospective longitudinal field investigation with an embedded prequential controlled computational experiment in a large general hospital. In practical terms, the AI-based digital twin (AI-DT) was followed over time while it encountered real indoor-air and other indoor-environmental problems arising during normal hospital operation. A prequential approach means that the AI-DT was tested on each new problem before it was allowed to learn from that problem. The hospital therefore provided a complex real-world setting in which the AI-DT could encounter different problems under changing environmental, spatial, temporal and operational conditions. The hospital itself was not the subject being investigated, and the study was not intended to determine how common indoor-environmental problems are across hospitals. The focus was on what the AI-DT could progressively learn to do in a real healthcare environment.

The methodology combined continuous environmental monitoring, targeted field investigation of problems, hospital building-system and operational information, occupants' perceived-environment information, independent verification of underlying causes, and prospective machine learning. Whenever an eligible problem occurred, the AI-DT received only the information that was genuinely available at that time. It then framed the problem, considered possible causes, identified any additional information needed and proposed possible solutions. These outputs were recorded and locked before the actual cause and outcome were known. Researchers then investigated the problem physically in the hospital, and independent reviewers determined, from the available evidence, what had most likely occurred and why. Only after this verification could the newly gained experience contribute to further AI-DT learning. This prevented the AI-DT from using future information to make an earlier diagnosis appear more accurate.

To determine whether improvement was genuinely due to learning, the progressively learning AI-DT was compared with a prespecified frozen non-learning digital-twin baseline. Both started from the same initial computational representation and received the same information available at the same time, but only the learning AI-DT was allowed to update what it had learnt from previous verified problems. The AI-DT also operated in shadow mode, meaning that its diagnoses and proposed solutions were recorded for research but were not used to direct routine hospital decisions. Hospital professionals continued their normal work, without any research intervention deliberately developing their professional learning. The study therefore determined what the AI-DT progressively learnt, whether that learning improved value-oriented problem diagnosis and solving beyond the non-learning baseline, and under what problem, environmental, spatial, temporal and operational conditions its capability remained reliable or reached its limits.

Machine-Learning Framework

The AI-DT needed a clearly defined process through which it could receive information, work with that information, learn from verified problem episodes and progressively update what it computationally represented about the hospital. This learning process was represented as:

$$Q_m = f(I_m, R_t)$$

$$A_m = f(Q_m)$$

$$L_m = f(A_m)$$

$$\text{Exp}_m = I_m \times L_m$$

$$R_{t+1} = f(R_t, \text{Exp}_m)$$

where I_m represented information available to the AI-DT from any source, including environmental measurements, hospital operational information and information provided by

humans; R_t represented the AI-DT's existing computational representation at time t ; Q_m represented machine-generated questions, diagnostic queries or problem propositions; A_m represented the algorithmic or computational processing used to examine those questions; L_m represented the resulting machine learning or computational refinement; Exp_m represented machine experience; and R_{t+1} represented the updated computational representation.

The complete learning process was therefore:

$$(I_m, R_t) \rightarrow Q_m \rightarrow A_m \rightarrow L_m \rightarrow Exp_m \rightarrow R_{t+1}$$

and its progressive development across successive verified hospital problems was represented as:

$$R_t \rightarrow R_{t+1} \rightarrow R_{t+2} \rightarrow \dots$$

In practical terms, the AI-DT received new information I_m within the context of what it already represented about the hospital R_t . It used these together to identify what needed to be examined Q_m , computationally processed the problem A_m , and learnt or refined its existing representation L_m . The interaction between the information received and what was learnt produced machine experience (Exp_m), which could then contribute to an updated computational representation R_{t+1} . Importantly, an update was permitted only after the relevant field problem had been independently investigated and verified, consistent with the test-before-learning design described in the methodology overview.

These equations were process equations, not statistical regression equations. In particular, the multiplication sign in $Exp_m = I_m \times L_m$ represented the interaction between information and machine learning rather than simple numerical multiplication. Likewise, machine experience did not imply consciousness or human-like experience. It referred specifically to the computational experience accumulated when new information was processed, learnt from and subsequently used to refine the AI-DT.

To make the learning process scientifically traceable, every stage was recorded. I_m was retained as timestamped input information; Q_m as the questions, diagnostic propositions or requests for additional information generated by the AI-DT; and A_m as the corresponding computational operations. L_m was evidenced by documented changes in model relationships, parameters, conditional probabilities, classifications or decision rules. Each R_t was preserved as an immutable version of the AI-DT. This made it possible to reconstruct what information the AI-DT had, how it processed that information, what it learnt, and how its computational representation changed from one verified problem episode to the next, rather than assuming that improved later performance necessarily demonstrated learning.

AI-DT Computational Architecture

Building on the machine-learning framework above, the AI-DT required a computational structure that could represent not only environmental data, but also how the hospital was physically organised, how environmental conditions changed over time, what different causes could explain a problem, and what had been learnt from previously verified problem episodes. For this reason, the AI-DT used a hybrid computational architecture rather than relying only on data-driven prediction or only on a fixed engineering model. The architecture combined a hospital spatial and building-system model, a dynamic environmental-state model, a probabilistic diagnostic model, and a problem-solution-experience model. These computational components were connected to a physical data-acquisition layer comprising environmental sensors, building-management-system data, operational records and other problem-specific information sources. This layer provided the real-time and historical information that formed part of I_m and connected the physical hospital environment to its evolving digital representation.

The physical data-acquisition layer included fixed and portable sensors selected according to the environmental problems being investigated. Core continuous sensors measured variables such as CO_2 , $PM_{2.5}$, PM_{10} , air temperature and relative humidity, while problem-dependent instrumentation measured variables such as airflow, air velocity, differential pressure, volatile organic compounds, selected gases, acoustic conditions and lighting conditions. Biological pollutants were assessed using targeted biological sampling where relevant, while perceived air quality, including odour, freshness, stuffiness and air acceptability, was obtained from appropriately structured occupant information rather than from physical sensors alone. Building-management-system data supplied additional information on variables such as fan operation, damper position, airflow, pressure control, filter condition and system alarms. Together, these sources allowed the AI-DT to receive information about both environmental conditions and the building and operational processes that could explain them.

The hospital spatial and building-system model represented the physical and engineering relationships relevant to problem diagnosis. These included the locations and functions of spaces, connections between adjoining areas, supply and return-air pathways, ventilation systems, pressure relationships, filtration arrangements and relevant equipment. The dynamic environmental-state model represented how indoor-air and other indoor-environmental conditions were expected to behave under changing occupancy, outdoor conditions, hospital activities and building-system operation. The probabilistic diagnostic model allowed the AI-DT to generate and compare competing explanations for an observed problem rather than automatically accepting the first plausible cause. The problem-solution-experience model retained verified relationships among previous problems, diagnostic evidence, interventions and outcomes so that experience from earlier episodes could contribute to later problem diagnosis and solving.

The sensors themselves did not constitute the AI-DT's understanding of the hospital. They produced measurements that became part of I_m only after appropriate timestamping, quality assurance, calibration checking and alignment with spatial, operational and contextual information. This distinction was important because a sensor measurement represented an observation of a physical condition, whereas the computational architecture was responsible for interpreting that observation in relation to the existing representation R_t , other available information and the problem being investigated. The computational architecture remained constrained by the physical reality of the hospital. Data-driven learning could refine relationships, parameters and diagnostic probabilities, but it could not create physically impossible spatial or HVAC connections. For example, the AI-DT could learn that a particular operational condition was repeatedly associated with poorer air quality in a neighbouring space, but such a relationship had to remain consistent with the actual ventilation and spatial connections within the hospital. Sensor location was therefore linked explicitly to the spatial and building-system model. Each fixed sensor was associated with a known physical location, space, ventilation zone and relevant building-system relationships. This allowed the AI-DT to interpret a measurement not simply as a numerical value, but as an observation occurring at a particular place, time and operating condition. Portable and targeted measurements could subsequently be added when the fixed monitoring network did not provide sufficient spatial or diagnostic resolution.

When a new problem emerged, the AI-DT first compared the observed condition with the expected or intended condition. It then generated competing explanations for the difference and assessed the evidence supporting or contradicting each explanation. Where the available I_m was insufficient to distinguish reliably among possible causes, Q_m identified what additional information would be most useful for reducing diagnostic uncertainty. This could lead, for example, to a request for additional airflow measurements, occupancy information, filter-condition data, biological sampling, perceived-air-quality information or another problem-specific investigation. This meant that sensing was not limited to passive continuous monitoring. The AI-DT could also identify when the existing sensor network or available information was insufficient and specify the additional measurement needed to discriminate among competing explanations. The resulting targeted field measurement then became new information only after it had actually been collected and quality checked. In this way, Q_m could influence what was measured next without allowing the AI-DT to invent information that had not been observed.

Importantly, the AI-DT was not required to produce a definitive diagnosis when the available evidence did not justify one. It could explicitly conclude that there was insufficient information for reliable diagnosis and identify what information was still required. This ability to recognise uncertainty was treated as part of competent value-oriented problem diagnosis and solving, because an unjustifiably confident diagnosis could lead to inappropriate interventions, wasted resources or unintended consequences. Conceptually, the complete cyber-physical information pathway was therefore: Physical hospital condition

→ sensors and other information sources → quality-controlled (I_m, R_t) → Q_m → A_m → L_m → Exp_m → R_{t+1} . The sensor and data-acquisition layer therefore served as the empirical interface between the physical hospital and the AI-DT, while the computational architecture determined how those observations were organised, interpreted, questioned, learnt from and incorporated into the evolving representation of the hospital.

Research Context, Duration and Baseline Characterisation

After defining how the AI-DT received and processed information, the next step was to establish the real hospital environment in which it would learn. The study was carried out in one continuously operating large general hospital with mechanically ventilated clinical and non-clinical spaces, an operational building-management system and access to environmental, maintenance and operational records. The hospital was selected because its changing occupancy, activities, space use and building operation could produce recurring indoor-environmental problems under different real-life conditions. The field study lasted 12 months and was preceded by an 8-week commissioning and baseline-characterisation phase. During the 8-week commissioning and baseline-characterisation phase, the sensors and data connections were installed, tested and calibrated. The relevant hospital spaces, ventilation systems and building-system relationships were also mapped and verified. Measurements collected during this period established the baseline environmental and operational conditions before prospective AI-DT learning began.

The initial computational representation, R_0 , was then constructed from all relevant pre-study information. This included the hospital's physical and engineering information, historical records from the previous 12 months, and the verified baseline measurements and field observations collected during the 8-week commissioning period. In simple terms, R_0 represented what the AI-DT computationally knew about the hospital before prospective learning started, including how spaces and systems were connected, how they normally operated, the environmental conditions normally observed, and the recurring problems previously recorded. Prospective learning was then assessed from this fixed starting point: $R_0 \rightarrow R_1 \rightarrow R_2 \rightarrow R_3 \rightarrow \dots$

The historical and 8-week baseline information therefore helped construct R_0 but did not count as evidence of prospective learning. Only new verified problem episodes occurring during the 12-month prospective study could contribute to learning from R_0 onwards.

Indoor-Air and Other Indoor-Environmental Measurements

After establishing R_0 , the prospective field programme continuously measured the hospital conditions that could contribute to indoor-environmental problems and provide information for I_m . The measurements covered chemical, particulate, biological and perceived indoor-air quality, together with ventilation, pressure, thermal, acoustic and lighting conditions. The exact information included in I_m depended on the problem being investigated rather than

being identical for every episode. Core continuous measurements included CO₂, PM_{2.5}, PM₁₀, air temperature and relative humidity. CO₂ was measured using non-dispersive infrared instruments, while PM_{2.5} and PM₁₀ were measured using optical particle instruments checked against appropriate reference or gravimetric methods. Where a problem suggested a chemical source, additional measurements included total volatile organic compounds, selected individual VOCs, formaldehyde, CO, NO₂, O₃ and other relevant gases. Total-VOC measurements were not used as a substitute for identifying specific compounds when required.

Ventilation measurements included air velocity, supply and return airflow where accessible, differential pressure and relevant building-management-system variables. Thermal measurements included air temperature and relative humidity, with mean radiant or operative temperature added where necessary. Acoustic conditions were assessed using appropriate sound-level measures, while lighting was assessed using horizontal and vertical illuminance and other relevant visual or biologically important measures where required. Continuous measurements were collected at 1-minute intervals where technically feasible. Raw data were retained, while five-minute summaries were generated for routine modelling using the median, minimum, maximum and short-term variability. The original 1-minute data remained available when brief events required closer investigation.

These measurements formed only part of the information available for diagnosis. Depending on the problem, I_m could also include walk-through observations, structured discussions or interviews with occupants and hospital personnel, complaints, maintenance and cleaning records, occupancy and activity information, equipment and operational records, building-management-system data, photographs and other problem-specific investigations. These additional information sources were integrated with measured environmental data and are described in the subsequent sections on contextual information and field diagnostic investigation.

Biological Pollutants and Perceived Air Quality

Where a problem suggested a possible biological cause, the investigation was extended beyond routine sensor measurements to include biological pollutants. These included airborne bacteria and fungi, including mould, and other microorganisms or biological material relevant to the problem. Biological investigation was triggered by evidence such as moisture or water intrusion, filtration failure, visible or suspected microbial growth, unusual odour, relevant occupant complaints or hospital activities that could reasonably introduce or spread biological contaminants. Where airborne microorganisms were relevant, active volumetric air sampling was carried out using calibrated microbial air samplers. Results were expressed as colony-forming units per cubic metre of air. Sampling volume, culture medium, incubation conditions and laboratory procedures were predefined. Where necessary, indoor results were compared with outdoor air or suitable comparison

spaces to help determine whether microorganisms were being amplified within the building rather than simply reflecting normal background levels.

Surface, filter or material samples were also collected where a local source or reservoir was suspected. Quantitative polymerase chain reaction (qPCR) or other targeted molecular methods could be used when culture-based methods alone could not answer the diagnostic question. The presence of microorganisms by itself was not treated as proof of a biological indoor-air problem. Their concentration, location, environmental conditions and likely source were considered together. Perceived air quality was also assessed because instruments alone could not fully describe how occupants experienced the air. Short standardised assessments recorded air acceptability, freshness or stuffiness, odour intensity and, where relevant, odour character. Responses were linked to the time and location of corresponding measurements. Poor perceived air quality did not automatically mean toxic exposure, while absence of odour did not prove that the air was healthy. Perceived and measured information were therefore treated as complementary evidence contributing to I_m .

Measurement Quality Assurance and Contextual Information

Because the AI-DT would learn from the information supplied to it, the study first had to ensure that this information was reliable, correctly timed and interpreted in the right context. Otherwise, an apparent AI-DT error could simply be the result of a faulty sensor, incorrect timing or missing information. All measuring instruments were therefore fully documented, including their manufacturer, model, accuracy, calibration history and deployment location. Temperature sensors were required to measure within about $\pm 0.3^\circ\text{C}$, and relative-humidity sensors within about $\pm 2\%$ RH over the relevant hospital range. Before installation, sensors were placed together and checked to confirm that they gave comparable readings. These checks were repeated at least every three months, after relocation or suspected malfunction, and at the end of the study.

All sensors and hospital records were also synchronised to the same time reference. This was important because a change in air quality, for example, might need to be linked accurately to an activity, a door opening, a change in ventilation operation or a maintenance event occurring at the same time. Automated checks were used to identify clearly impossible readings, duplicated records, timing errors, sensors that appeared to become stuck, and sudden changes that could not reasonably have occurred. Unusually high or low readings were not automatically deleted, because they might represent a genuine environmental problem. Occasionally, a sensor could fail briefly and leave a small gap in the record. A gap of 5 minutes or less could be estimated only in secondary analyses when the readings immediately before and after the gap showed that conditions had remained stable. Longer gaps were left as missing rather than filled with assumed values. This prevented invented data from influencing the main findings.

The AI-DT was also assessed using only the information that was actually available when it made its decision. Information discovered later during field investigation could be used to determine whether the AI-DT had been correct, but could not be inserted afterwards into the earlier I_m . This prevented hindsight bias and ensured a fair test of what the AI-DT could genuinely diagnose at that time. Environmental measurements were interpreted together with relevant contextual information, such as occupancy, activities, cleaning, maintenance, ventilation-system operation, equipment use, complaints, interviews and walk-through observations. Therefore, I_m represented the reliable and auditable information genuinely available to the AI-DT at each decision point.

Definition of Problem Episodes

After establishing how reliable information was collected, the next step was to define what counted as a problem. The study did not treat every unusual sensor reading as a separate problem. Instead, the main unit of analysis was a problem episode. A problem episode was a specific event, occurring within a particular place and period of time, in which the actual indoor-environmental condition was meaningfully different from the condition required for occupants, activities or hospital operations to function as intended. Therefore, a high pollutant reading or another threshold exceedance alone did not fully define a problem. The study also considered what condition was expected, what was actually happening, who or what was affected, and why the difference mattered.

Rules for deciding when an episode started and ended were defined in advance for each type of problem. If conditions returned fully to the expected state and the same problem later occurred again, this was treated as a new episode. However, several abnormal readings caused by the same unresolved underlying problem were treated as one episode rather than several. This prevented the same problem from being counted repeatedly. A recurring problem was defined as the same, or a closely related, problem occurring on at least three separate occasions. Between four and six principal problem families were selected before prospective evaluation. These covered indoor-air problems and at least two other indoor-environmental domains, such as thermal, acoustic or lighting conditions. The research observed problems that arose naturally during hospital operation. No unhealthy or hazardous condition was deliberately created.

How the AI-DT Operated in Practice

In practical terms, the AI-DT was a software-based computational representation of the relevant hospital spaces, building systems and indoor environment. It was not a physical robot or a separate physical copy of the hospital. It operated on a computing platform and received information collected from the real hospital, including environmental sensor measurements, relevant building-management-system data, occupancy and activity information, maintenance and operational records, and appropriately recorded information from occupants and hospital personnel. The AI-DT connected this information to its

representation of the hospital's rooms, ventilation pathways, equipment, pressure relationships and other relevant physical and operational relationships. This allowed it to examine a problem occurring at a particular place and time rather than simply responding to isolated sensor readings. Its outputs included how it framed the problem, possible causes, its level of uncertainty, additional information it considered necessary, possible solutions and their expected value. The overall procedure could therefore be visualised simply as:

Real hospital problem → information collected → AI-DT examines the problem → AI-DT assessment is locked → researchers investigate the problem physically → actual cause is independently established → AI-DT assessment is compared with what actually happened → verified experience contributes to subsequent learning

The AI-DT therefore generated the prospective diagnosis, while the subsequent physical field investigation provided the evidence needed to determine whether that diagnosis was correct.

Prospective Field Experimental Procedure

Each eligible problem episode followed the same problem-diagnosis-verification-learning cycle. When continuous monitoring, occupant reports, building-management-system information or routine hospital activities indicated a possible problem, the episode was formally registered before its underlying cause was known. The location, time and all relevant information genuinely available at that moment were recorded. This episode-specific information constituted I_m . The same I_m was then given independently to two computational systems. The first was the progressively learning AI-DT, which could use what it had learnt from earlier verified problem episodes. The second was the frozen non-learning digital twin, which remained at its initial computational representation R_0 and could not learn from previous episodes.

Both systems were asked to examine the same unresolved real-world problem. Each produced its problem framing, identified abnormalities, generated possible causes, expressed uncertainty, identified any additional information required, proposed possible solutions and considered their likely value. Their outputs were timestamped and locked before the actual cause and outcome became known. This created a fair comparison because the progressively learning AI-DT and the frozen non-learning digital twin received exactly the same information about the same problem at the same time. The important difference was that only the learning AI-DT could benefit from previous verified experience. After both systems had made their assessments, the research team investigated the problem physically in the hospital. Depending on the possible causes, this could include portable pollutant measurements, measurements across different locations, airflow and pressure testing, inspection of fans, dampers and filters, chemical or microbial sampling, moisture investigation, thermal, acoustic or lighting measurements, equipment inspection,

walk-through observations, and review of occupancy, activities, cleaning and maintenance records.

Information discovered during this later investigation was used to determine what had actually happened. It could not be added retrospectively to the earlier I_m , because the purpose was to judge what each system could diagnose using only the information that was genuinely available at the time. Where hospital personnel authorised corrective action, conditions were monitored before, during and after the intervention. The basic reasoning was:

Measured problem → possible explanations → targeted investigation → identified cause → authorised corrective action → predicted response → observed response

If correcting the suspected cause produced the expected improvement, confidence in the diagnosis increased. If the expected improvement did not occur, the diagnosis had to be reconsidered.

Independent Reference Diagnosis

After each episode had been sufficiently investigated, all relevant evidence was brought together into an independent reference package. This could include environmental measurements, building-system information, field observations, equipment inspections, maintenance findings, occupancy and activity information, biological or chemical analyses, perceived-air-quality information and the response to any corrective action. A three-member independent reference panel then reviewed the evidence. The panel collectively represented expertise in HVAC or building services, indoor environmental quality and hospital facilities or occupational/environmental health. Members first assessed each episode independently before discussing the evidence and reaching a final consensus.

A cause was classified as verified when strong evidence established what had happened, for example through direct measurement, inspection or a corrective action followed by the predicted environmental response. A cause was classified as probable when the available evidence supported the explanation but could not establish it conclusively. Agreement among panel members was statistically assessed before consensus. Episodes without a defensible verified cause were not discarded. They remained useful for examining whether the AI-DT could recognise that something was abnormal or that available information was insufficient. However, such episodes were excluded from analyses that required a confirmed underlying cause.

Progressive Learning, Frozen Baseline and Human-Learning Condition

The next step was to determine whether the AI-DT genuinely improved through learning. The progressively learning AI-DT was represented as $P_{DT,l}$, while the frozen non-learning digital twin was represented as $P_{DT,o}$. Both started from the same initial computational

representation, R_0 , and received the same information available at the same time. The frozen system could still analyse each new problem, but it could not update what it had learnt from previous problems. Its pathway therefore remained:

$$R_0 \rightarrow R_0$$

The learning AI-DT, in contrast, could progressively update its computational representation:

$$R_0 \rightarrow R_1 \rightarrow R_2 \rightarrow \dots \rightarrow R_n$$

Only problem episodes that had been independently investigated and adjudicated were allowed to contribute to later machine learning. Updates occurred at fixed four-week intervals, with each update producing a new archived model state. The sequence could therefore be visualised as:

$$\text{Problem}_1 \rightarrow \text{test } R_0 \rightarrow \text{investigate} \rightarrow \text{verify} \rightarrow \text{Exp}_{m1} \rightarrow R_1 \rightarrow \text{later problem} \rightarrow \text{test } R_1 \rightarrow \text{investigate} \rightarrow \text{verify} \rightarrow \text{Exp}_{m2} \rightarrow R_2$$

The AI-DT operated in shadow mode throughout RQ1. Its diagnoses and proposed solutions were recorded for research but were not shown to the hospital professionals managing the corresponding problem. This was important because RQ1 aimed to examine AI-DT learning without deliberately developing professional human learning at the same time. Professionals continued their normal work and could still learn naturally through experience or external training, but the research did not introduce an intervention intended to produce $M_t \rightarrow M_{t+1}$. Any relevant external professional training was recorded.

Assessment of Value-Oriented Problem Diagnosis and Solving

AI-DT performance, represented as P_{DT} , was not judged using a single accuracy score. Instead, it was assessed across eight dimensions: problem framing and definition; abnormality recognition; distinguishing symptoms from causes; recognising spatial, temporal and operational relationships; generating and comparing possible causes; identifying additional information needed to reduce uncertainty; generating and assessing possible solutions; and evaluating the value of those solutions. A separate three-member evaluation panel, different from the panel that established the reference diagnosis, assessed anonymised outputs from $P_{DT,i}$ and $P_{DT,o}$ without knowing which system or model version had produced them. Three evaluators provided independent expert judgement while keeping repeated assessment of many problem episodes feasible within the PhD. Each evaluator scored every output independently using a predefined 0–4 anchored rubric, ranging from absent or seriously incorrect performance to strong performance consistent with the independently established evidence.

The rubric was first tested using historical problem episodes that were excluded from the prospective analysis. Any unclear scoring descriptions were refined before the main study.

Agreement among evaluators was then quantified using the intraclass correlation coefficient (ICC) with 95% confidence intervals, with $ICC \geq 0.75$ used as the predefined minimum level of acceptable reliability. Because the individual dimension scores were ordinal, weighted kappa was also used as a complementary measure of agreement for those ratings. Value orientation was assessed separately from technical correctness. The AI-DT had to consider who was affected, the undesirable current condition, the required condition or goal, why achieving that goal mattered, the expected usefulness of a proposed solution, the resources or sacrifices required, operational constraints and possible unintended consequences. A cheap or easy solution could not be judged as high value if it failed to solve the problem, while a technically effective solution could not be rated highly if it created unacceptable safety, infection-control or operational consequences. Technical measures, including sensitivity, specificity, precision, recall, false-positive rate, detection time, ranking of the verified cause and probabilistic calibration, were also examined. These remained secondary to the broader assessment of value-oriented problem diagnosis and solving.

Evaluation of AI-DT Learning, Performance and Capability Boundaries

The study examined not only whether the learning AI-DT performed better than the frozen system, but also where its capability began to weaken or fail. Capability boundaries were assessed across five domains: problem conditions such as diagnostic ambiguity; environmental conditions such as magnitude and variability; spatial conditions such as ventilation connectivity and source proximity; temporal conditions such as duration, recurrence and delayed effects; and operational conditions such as occupancy, activities, HVAC operation, cleaning and maintenance. A single failure was not treated as a boundary. A boundary required repeated deterioration under an identifiable condition, together with a practically meaningful reduction in performance or increase in diagnostic error, uncertainty or poor value-oriented solving. Comparisons with fewer than 10 evaluable episodes with and without the condition being examined were treated as exploratory. The study targeted at least 100 independently adjudicable problem episodes. Sample-size adequacy was assessed by statistical simulation using pilot estimates of problem frequency, performance variation and repeated episodes. The design targeted at least 90% statistical power at a two-sided $\alpha = 0.05$, meaning a high probability of detecting a practically important difference if one genuinely existed. Monitoring could be extended by up to three months if too few episodes occurred.

For each episode, the main comparison involved the learning AI-DT output ($P_{DT,i}$), the frozen digital-twin output ($P_{DT,o}$), and the independently adjudicated reference diagnosis. The reference diagnosis was established after physical investigation using measurements, inspections, building-system and maintenance information, chemical or biological analyses where required, and the response to authorised corrective action. The independent reference panel then classified the cause as verified, probable or unresolved. Both outputs were judged against the same real-world evidence. Because several episodes could occur

in the same hospital location or belong to the same problem family, they were not treated as completely independent. Mixed-effects models were therefore used. In simple terms, these models compared the two systems while recognising that episodes from the same place or problem type might naturally resemble one another. Ordinal outcomes such as 0–4 performance scores, binary outcomes such as correct or incorrect cause identification, and continuous outcomes such as detection time were analysed using corresponding mixed-effects models.

The overall learning-versus-frozen comparison was tested first, followed by individual performance dimensions. Because multiple tests increase the chance of apparently significant findings occurring by chance, Holm adjustment controlled this risk. Results were reported using 95% confidence intervals, effect sizes and exact P values. Statistical significance alone was insufficient; improvement also had to be large enough to matter in practice and could not involve unacceptable deterioration in a safety-critical dimension. Sensitivity analyses examined whether conclusions changed after excluding episodes with substantial missing data, uncertain reference diagnoses or confirmed sensor malfunction. Ablation analyses removed information classes from I_m , one at a time, to determine which sources were important for reliable diagnosis. Every missed problem, incorrect high-confidence diagnosis and potentially unsafe recommendation underwent failure analysis and was retained as evidence of capability boundaries.

The study also retained a computational audit trail showing how learning occurred:

$$(I_m, R_t) \rightarrow Q_m \rightarrow A_m \rightarrow L_m \rightarrow \text{Exp}_m \rightarrow R_{t+1}$$

This allowed the study to examine both whether performance improved and how the AI-DT's computational representation changed from R_t to R_{t+1} . Every computational state and software version was preserved so the learning pathway could be reconstructed. Monitoring, sampling, scoring, adjudication and statistical procedures were fixed before the final locked evaluation, during which the evaluation rules and AI-DT architecture could not be changed. Safety was built into the study design. The AI-DT operated only as a decision-support research system and did not autonomously control hospital systems. Any corrective action used for verification was carried out through hospital procedures by authorised personnel. The research team did not deliberately create hazardous chemical exposure, microbial contamination, thermal stress, excessive noise or other unhealthy conditions. Human-related information was collected only at the level necessary and handled under appropriate ethical and data-governance requirements.

The null hypothesis was:

$$H_{01}: P_{DT,i} \leq P_{DT,0}$$

where $P_{DT,i}$ represented value-oriented problem-diagnosis-and-solving performance of the progressively learning AI-DT and $P_{DT,0}$ represented performance of the frozen non-

learning baseline. The alternative hypothesis was:

$$H_{11}: P_{DT,I} > P_{DT,0}$$

for one or more prespecified dimensions and conditions. Because P_{DT} was multidimensional, H_{11} did not require superiority for every problem under every condition. Support for H_{11} required improvement that was reproducible, statistically supported and practically meaningful, followed by identification of the conditions under which improvement persisted or deteriorated. The final scientific output was a capability-boundary map of progressive AI-DT learning. It showed which dimensions of value-oriented problem framing, diagnosis and solving improved as $R_t \rightarrow R_{t+1} \rightarrow R_{t+2} \rightarrow \dots$, which capabilities plateaued or failed to improve, which information streams were necessary for reliable diagnosis, and under which problem, environmental, spatial, temporal and operational conditions the AI-DT reached its limits.

The defining sequence was: real unresolved hospital problem $\rightarrow$ prospective AI-DT diagnosis $\rightarrow$ locked output $\rightarrow$ physical investigation $\rightarrow$ independent adjudication $\rightarrow$ verified experience $\rightarrow$ AI-DT learning $\rightarrow$ later real-world challenge. This provided the empirical basis for determining what progressive AI-DT learning could achieve without deliberately developing corresponding professional human learning.

Ethical Considerations and Methodology Contribution to Knowledge

The study was designed to investigate AI-DT learning without introducing additional risk to hospital occupants or operations. The AI-DT operated in shadow mode and did not autonomously control ventilation, pressure, filtration, temperature or other hospital systems. Any corrective action used to investigate or verify a problem was undertaken only through established hospital procedures by authorised personnel. The research did not deliberately create hazardous chemical exposure, microbial contamination, thermal stress, excessive noise or other unhealthy indoor conditions. Human-generated information, including complaints, interviews and perceived-air-quality responses, was collected only to the extent necessary for the research and handled in accordance with institutional ethics, informed-consent, privacy and data-governance requirements. Personally identifiable information was minimised, while access to hospital operational and building-system information was controlled appropriately.

The principal methodological contribution to knowledge was the development of a real-world framework for determining whether an AI-DT genuinely learns to diagnose and support solving problems over time, rather than merely demonstrating predictive accuracy on an existing dataset. The methodology combined a prospective longitudinal field investigation with an embedded prequential controlled computational experiment. The progressively learning AI-DT and frozen non-learning digital twin began from the same R_0 , received the same contemporaneous I_m , and produced locked assessments before the actual cause of each problem was known. Subsequent physical investigation and

independent adjudication provided the reference against which both systems were evaluated. This approach also made the learning pathway auditable through:

$$(I_m, R_t) \rightarrow Q_m \rightarrow A_m \rightarrow L_m \rightarrow \text{Exp}_m \rightarrow R_{t+1}$$

The methodology therefore provides a reproducible way to distinguish learning from apparent improvement, examine value-oriented problem diagnosis and solving, and identify the problem, environmental, spatial, temporal and operational conditions that define AI-DT capability boundaries.

Methods for Research Question 2:

Overview

The methodology for Research Question 2 was designed as a prospective longitudinal field intervention with an embedded prequential human–AI-DT experiment, conducted as a direct continuation of the methodology for RQ1 in the same large general hospital. The methodology for RQ1 had established what the AI-DT could progressively learn when professional human learning was not deliberately developed. The methodology for RQ2 retained the same hospital setting, monitoring system, problem definitions, field-investigation procedures, independent reference diagnosis and assessment of value-oriented problem diagnosis and solving. The main methodological change was the deliberate introduction of professional human learning and reciprocal human–AI-DT interaction. The methodology for RQ2 did not begin with a new AI-DT. The final prospectively evaluated AI-DT from RQ1 became the starting point. Hospital professionals involved in indoor-environmental problem solving were deliberately supported to improve how they questioned information, reasoned about possible causes, formed judgements, made decisions and reflected on verified outcomes. Their developing mental models were then allowed to interact with, challenge and contribute information to the continuously learning AI-DT.

For each eligible real hospital problem, the methodology distinguished three stages of reasoning before the actual cause was known: independent professional human reasoning, independent AI-DT reasoning, and final coupled human–AI-DT reasoning. Their outputs were locked before subsequent physical field investigation and independent adjudication established what had actually happened. Verified outcomes were then used for structured professional reflection and further AI-DT learning. Repeating this process allowed the methodology for RQ2 to determine whether deliberate professional human learning added capability beyond AI-DT learning alone, how that additional capability emerged, and under which conditions human–AI-DT complementarity strengthened, weakened or failed to emerge.

Human-Learning Framework

To examine how deliberate professional learning contributed to the methodology for RQ2, the human-learning process was represented as:

$$Q_h = f(I_h, M_t)$$

$$T_h = f(Q_h)$$

$$L_h = f(T_h)$$

$$\text{Exp}_h = I_h \times L_h$$

$$M_{t+1} = f(M_t, \text{Exp}_h)$$

where I_h represented information received by the professional from any source, including the AI-DT; M_t represented the professional's existing mental model at time t ; Q_h represented the questions generated from that information and existing understanding; T_h represented the thinking process, including critical thinking, reflective thinking, logical reasoning, abstract reasoning and creative imagination; L_h represented human learning; Exp_h represented human experience; and M_{t+1} represented the updated mental model.

The complete human-learning process was therefore:

$$(I_h, M_t) \rightarrow Q_h \rightarrow T_h \rightarrow L_h \rightarrow \text{Exp}_h \rightarrow M_{t+1}$$

These were process equations, not statistical regression equations. The multiplication sign in $\text{Exp}_h = I_h \times L_h$ represented the interaction between information and learning rather than numerical multiplication. Importantly, I_h represented information received by the professional from any source, including environmental measurements, field observations, other people, documents and records, and the AI-DT. Information produced by the AI-DT therefore formed only one possible component of I_h . Information from any source was not automatically treated as human knowledge or understanding; the professional still had to question, think about, interpret and learn from that information.

Internal thinking and mental models could not be observed directly. The methodology therefore examined their external evidence. Q_h was captured through the questions professionals asked. Evidence of T_h came from recorded reasoning showing how they challenged assumptions, compared explanations, connected evidence and considered alternatives. L_h was indicated by defensible improvement in knowledge, reasoning, judgement or decision quality after engagement with information and verified outcomes. Development from $M_t \rightarrow M_{t+1}$ was therefore inferred from repeated changes in how professionals externally represented their knowledge and understanding in the context of the problem at hand, taking into consideration causal relationships, problem structure, stakeholder requirements and value relationships. This allowed the methodology for RQ2 to examine professional learning without claiming direct measurement of unobservable internal cognition.

Reciprocal Human–AI-DT Learning Framework

Building on the human-learning framework above, the methodology for RQ2 examined reciprocal learning between professionals and the AI-DT. The purpose was not simply to place humans and an AI-DT together, but to investigate whether learning by the professional could contribute to subsequent machine learning and whether learning by the AI-DT could, in turn, contribute to further professional learning. As professionals received I_h and considered it in relation to their existing mental model M_t , this interaction could trigger Q_h . The questions generated then guided T_h , the thinking process through which professionals critically and reflectively examined the real hospital problem, compared possible explanations, and considered relationships, assumptions and alternatives. This thinking supported L_h , through which the information received could be refined into new or improved knowledge, understanding and skills. The interaction between the information received and the learning achieved contributed to Exp_h , representing the knowledge, understanding and skills gained through that learning experience. When this new experience was represented in the context of the problem at hand and integrated with existing experience, it could refine the professional's mental model from M_t to M_{t+1} .

The refined mental model M_{t+1} could then change how the professional viewed the problem and the information currently available to the AI-DT. For example, the professional might recognise that important information was missing, that an existing piece of information needed correction or clarification, or that an overlooked contextual relationship should be considered. The professional could therefore add to, correct or revise the information available to the AI-DT. Once this information became available to the machine, it formed new or revised I_m , because I_m represented information available to the AI-DT from any source, including humans. The pathway from professional learning to machine learning was therefore represented as: $M_{t+1} \rightarrow \text{human adds, corrects or revises information available to the AI-DT} \rightarrow \text{new or revised } I_m$.

The AI-DT then processed this new I_m in relation to its existing computational representation R_t :

$$(I_m, R_t) \rightarrow Q_m \rightarrow A_m \rightarrow L_m \rightarrow Exp_m \rightarrow R_{t+1}.$$

In this way, information arising from the professional's updated understanding could contribute to refinement of the AI-DT's computational representation from R_t to R_{t+1} . The professional did not directly alter R_t . Instead, the professional changed or added to the information available to the AI-DT, and the AI-DT then processed and learnt from that information through its own machine-learning pathway. The reciprocal process also operated in the opposite direction. The updated AI-DT could generate new analyses, questions, explanations or other information. When this information was received by the professional, it became part of I_h :

$$R_{t+1} \rightarrow \text{AI-DT-generated information} \rightarrow I_{h,t+1}$$

The professional could then consider this new information in relation to the existing mental model, potentially triggering another cycle of questioning, thinking, learning and mental-model refinement. Across repeated real hospital problem episodes, reciprocal development was represented as:

$$(I_h, M_t) \rightarrow \text{human learning} \rightarrow M_{t+1} \rightarrow \text{new/revised } I_m \rightarrow (I_m, R_t) \rightarrow \text{machine learning} \rightarrow R_{t+1} \\ \rightarrow \text{new } I_h \rightarrow (I_h, M_{t+1}) \rightarrow \text{further human learning} \rightarrow M_{t+2} \rightarrow \text{new/revised } I_m \rightarrow (I_m, R_{t+1}) \rightarrow \\ \text{further machine learning} \rightarrow R_{t+2} \rightarrow \dots$$

This representation showed that human and machine learning were reciprocally connected rather than necessarily occurring simultaneously or at the same rate. Learning by the professional could generate new or revised information for the AI-DT, while learning by the AI-DT could subsequently generate new information for further professional learning.

Information provided by a professional was not automatically treated as fact. For example, a professional might suggest that an unusual odour originated from a particular ventilation pathway. This could enter I_m as a human-generated hypothesis, but it remained labelled as a hypothesis until supported by measurement, inspection, field investigation or independent adjudication. This distinction prevented unsupported professional judgement from becoming machine ground truth.

The conceptual aim of this reciprocal relationship was complementarity rather than coexistence. Complementarity meant that deliberate interaction between the professional and AI-DT could produce better value-oriented problem diagnosis and solving than either could achieve independently on the same problem. The stronger form of complementarity was represented as:

$$P_{h+DT} > \max(P_h, P_{DT})$$

where P_h represented the professional's independent performance, P_{DT} represented the AI-DT's independent performance on the same problem, and P_{h+DT} represented their performance after deliberate interaction. This framework therefore defined what reciprocal development and complementarity meant conceptually; the later evaluation section specifies how these were measured and tested empirically.

Continuity From RQ1 and Parallel AI-DT Lineages

The methodology for RQ2 continued directly from RQ1 because the purpose was to determine what additional capability emerged when deliberate professional human learning was added to an already established learning AI-DT. Starting again with a new AI-DT would have made it difficult to know whether later differences were caused by professional human learning or simply by rebuilding the machine. Therefore, the final locked AI-DT representation from RQ1 became the common starting point for RQ2. The same monitoring network, problem definitions, reference-diagnosis procedure and performance

rubric were also retained so that findings from the two research questions remained directly comparable.

At the start of RQ2, this final RQ1 representation was duplicated to create two identical AI-DT lineages. This was necessary to create a fair comparison. Both systems began with exactly the same computational representation and previous machine experience, so any later difference could be linked more convincingly to what happened during RQ2. The AI-DT-only lineage continued learning as in RQ1 from independently verified environmental, operational and problem-outcome information, while the coupled human–AI-DT lineage, which began from the same computational state, received the same verified information plus any new or revised I_m arising from deliberately developed professional questioning, reasoning and updated mental models, with verification where required; running both lineages concurrently therefore provided a counterfactual comparison that helped isolate the additional contribution of deliberate professional human learning to AI-DT development from other possible influences such as season, changing hospital activities, new problem types or accumulating machine experience.

Professional Participants, Baseline Characterisation and Human-Learning Intervention

After establishing the two parallel AI-DT lineages, the methodology for RQ2 required professionals who were genuinely involved in solving indoor-air and other indoor-environmental problems. Participants therefore included hospital professionals whose normal roles involved framing, diagnosing or solving such problems, such as building-services and HVAC engineers, facilities-management professionals, indoor-environmental or occupational-health personnel and other relevant technical staff. Temporary personnel without meaningful diagnostic responsibility were excluded because the study aimed to examine development of professional problem-solving capability over time.

Approximately 24 professionals were targeted. This number provided a manageable group for repeated longitudinal assessment while allowing variation in professional roles and problem-solving experience. Final sample adequacy was checked using statistical simulation informed by the frequency and dependence of problem episodes observed during RQ1 and preliminary estimates of professional performance variation and repeated-measures dependence obtained during the initial RQ2 pilot phase. The same professionals were followed across multiple problem episodes because the methodology was concerned with how individual and team capability developed over time, rather than treating every decision as if it came from a different person.

The professional-learning phase was conducted over approximately 6–8 months. The first few weeks were used for baseline assessment and initial facilitated learning sessions. The remaining period focused mainly on episode-based learning embedded in real hospital problems, allowing the same professionals to encounter, reason through, reflect on and

learn from multiple problem episodes over time. This main longitudinal phase took place within the professionals' normal work setting and used naturally occurring indoor-air and other indoor-environmental problems encountered during routine hospital operation. No artificial environmental problems were created for training or evaluation. This duration provided sufficient opportunity to examine longitudinal development in professional learning and reciprocal human–AI-DT interaction while remaining feasible within the wider PhD programme and reusing the field infrastructure established in RQ1.

Before deliberate learning began, each participant completed a baseline assessment using previously adjudicated hospital cases that had not been used to train them. The purpose was to establish how they currently framed problems, generated questions, distinguished symptoms from underlying causes, identified missing information, proposed solutions and made value-oriented judgements. Participants also externalised their understanding of the relationships within each problem. This was done through a structured case-based assessment in which participants recorded their problem framing, questions, reasoning, judgements and proposed solutions, and produced a simple causal or problem-relationship map showing how they believed the important factors were connected.

Where clarification was required, a short semi-structured interview was conducted. This provided observable baseline evidence of M_t , while recognising that the internal mental model itself could not be measured directly. These previously adjudicated cases were used only to provide comparable baseline conditions and support the initial facilitated learning. The principal evidence of development from $M_t \rightarrow M_{t+1} \rightarrow M_{t+2}$ was generated from repeated engagement with real hospital problems, real operational constraints and real problem outcomes during normal professional practice. The same core dimensions assessed at baseline were then examined repeatedly during these real problem episodes so that changes in questioning, reasoning, judgement, decision-making, learning and externalised mental models could be tracked over time.

The deliberate professional human-learning intervention then focused on developing cognitive governance for value-oriented problem diagnosis and solving, meaning the professionals' capability to deliberately govern how they questioned information, thought through problems, formed judgements, made decisions and learnt from outcomes, rather than simply teaching additional factual content. Participants were supported to understand how received information I_h , when considered in relation to their existing M_t , could trigger questioning Q_h , which then guided thinking T_h , learning L_h , experience Exp_h and subsequent refinement towards M_{t+1} . The intervention developed the practical use of questioning, critical and reflective thinking, logical reasoning, abstract reasoning and creative imagination.

Professionals learnt to distinguish symptoms from causes, challenge assumptions, compare competing explanations and interrogate AI-DT outputs rather than automatically accepting them. For each problem, participants were required to establish the current

condition, required condition, gap, affected stakeholders, goal and significance of achieving that goal. They then generated questions to identify missing information and discriminate among possible causes. Proposed solutions were assessed not only for technical effectiveness, but also for expected usefulness, invested resources, operational constraints and possible unintended consequences.

Initial facilitated learning was followed by episode-based learning during real hospital problems. After each problem had been independently investigated and adjudicated, professionals received structured feedback comparing their original questions, reasoning, judgements and decisions with what was later established. They then reflected on what assumptions were supported or contradicted, what information had been missed, what they had learnt and how that experience should influence future problems. The intervention protocol, facilitation prompts and reflection procedures were standardised before formal evaluation. Importantly, the intervention developed how professionals learnt and reasoned; it did not provide advance answers to later hospital problems.

Field Platform and Environmental Information

The methodology for RQ2 continued to use the same field-monitoring platform established in RQ1. The same hospital monitoring systems, sensors, problem definitions, data-quality procedures, field-investigation methods and independent reference-diagnosis procedures were retained. This was important because any difference observed during RQ2 should reflect the introduction of deliberate professional human learning and human–AI-DT interaction, rather than a change in how environmental problems were measured or investigated.

For each problem episode, the available information could include chemical, particulate, biological and perceived indoor-air-quality information, together with thermal, acoustic, lighting, airflow, pressure and building-system conditions. Relevant contextual information could also include occupancy, space use, clinical and non-clinical activities, cleaning, maintenance, outdoor conditions, complaints, previous interventions and their outcomes. Not every type of information was automatically required for every problem. Identifying which information was relevant, what was missing and what additional information was needed to reduce uncertainty remained part of the problem-diagnosis process.

Prospective Human–AI-DT Field Procedure

Each eligible recurring hospital problem was registered before its underlying cause was known. At that point, all relevant environmental, spatial, temporal, operational and human-generated information genuinely available was recorded. Information discovered later during investigation was kept separate so that the professional and AI-DT could be evaluated using only what was actually known when they made their initial assessments. The responsible professional first examined the unresolved problem without seeing the AI-DT's assessment. The professional recorded the problem as understood at that time, the

questions that needed to be asked, possible causes, important missing information, uncertainty, and an initial judgement or proposed action where sufficient evidence existed. This independent professional assessment constituted P_h .

At the same time, the AI-DT-only lineage independently examined the same unresolved problem using the contemporaneously available I_m . It recorded its problem framing, possible causes, uncertainty, additional information required and proposed solutions. This independent AI-DT assessment constituted P_{DT} . Both P_h and P_{DT} were therefore produced independently for the same problem before either could influence the other. The professional then interacted with the coupled human–AI-DT lineage. During this interaction, the professional could question the AI-DT’s reasoning, request supporting evidence, challenge assumptions, suggest alternative explanations, identify overlooked contextual information and request additional measurements. The professional could also add, correct or clarify information available to the coupled AI-DT. Human-generated hypotheses remained clearly identified as hypotheses until appropriate evidence supported them.

The interaction was recorded so that important questions, responses, newly introduced information and changes in judgement or confidence could later be reconstructed. After the interaction, the professional and coupled AI-DT produced a final problem framing, causal judgement, uncertainty assessment, request for further investigation where necessary, and proposed solution or alternatives. This final coupled assessment constituted P_{h+DT} . The methodology therefore produced three directly comparable assessments for the same unresolved real hospital problem: the professional’s independent assessment (P_h), the AI-DT’s independent assessment (P_{DT}), and the final coupled professional–AI-DT assessment (P_{h+DT}) after deliberate interaction, allowing the study to determine how the professional and AI-DT performed independently, what changed through their interaction, and whether the coupled performance exceeded either one acting independently.

Field Verification, Feedback and Reciprocal Learning

After the three assessments for each problem episode had been completed, P_h , P_{DT} and P_{h+DT} were locked and could no longer be changed. The research team then investigated the actual problem in the relevant hospital location to establish, as objectively as possible, what had happened and why. Depending on the problem, this could involve portable environmental measurements, airflow and pressure testing, equipment inspection, chemical or biological sampling, moisture investigation, review of occupancy and activities, and observation of any authorised corrective action. Information obtained during this stage was used for verification and was not added retrospectively to the earlier assessments. The evidence generated from the physical investigation was then reviewed by the independent reference-diagnosis panel. The purpose of this panel was not to judge whether the professional or AI-DT had performed well. Its role was to establish the best-supported real-world explanation of the problem and classify the underlying cause as

verified, probable or unresolved. The resulting reference diagnosis therefore represented the common real-world evidence against which the three assessments made before the cause was known could subsequently be evaluated.

A separate evaluation panel then assessed anonymised versions of P_h , P_{DT} and P_{h+DT} for the same problem episode. Using the predefined dimensions of value-oriented problem diagnosis and solving, the panel compared each assessment with the independently adjudicated reference evidence. In practical terms, the comparison asked how well the professional acting alone (P_h) understood and addressed the problem, how well the AI-DT acting alone (P_{DT}) performed, and whether the professional and AI-DT working together (P_{h+DT}) produced a better assessment than either achieved independently. The main comparison for answering RQ2 was P_{h+DT} versus P_{DT} , because this showed whether adding deliberate professional human learning and interaction produced capability beyond AI-DT learning alone. A stronger test of complementarity examined whether $P_{h+DT} > \max(P_h, P_{DT})$. This meant that genuine complementarity was demonstrated when the coupled professional–AI-DT performance exceeded both the professional acting independently and the AI-DT acting independently on the same unresolved real hospital problem.

Only after the problem had been physically investigated and independently adjudicated was the episode used for further learning. Professionals received structured feedback showing how their original independent assessment, the AI-DT's independent assessment and the final coupled assessment compared with the established evidence. They then reflected on what they had understood correctly, what had been missed, which assumptions were unsupported and what knowledge or understanding should influence their approach to later problems. This new learning experience contributed to Exp_h and could refine the professional's mental model from M_t to M_{t+1} .

The two AI-DT lineages were also updated separately after adjudication. The AI-DT-only lineage continued to learn from the independently adjudicated environmental and operational evidence in the same way established in RQ1. The coupled human–AI-DT lineage received the same verified evidence, together with any appropriately verified new information or contextual relationships that had emerged through professional human–AI-DT interaction. Their later performance could therefore show whether information arising from deliberately developed professional learning contributed additional capability to subsequent machine learning.

The overall procedure therefore moved from an unresolved real hospital problem, through independent professional and AI-DT assessment, deliberate human–AI-DT interaction and coupled assessment, to physical field investigation and independent reference diagnosis. The three locked assessments were then evaluated against the same real-world evidence, after which professional reflection and separate updating of the two AI-DT lineages occurred before the next problem episode. This sequence made it possible to test

reciprocal human–AI-DT learning prospectively across successive real hospital problems, rather than judging its value merely from the apparent quality of individual human–AI interactions.

Measurement of Human Learning and Mental-Model Development

Having established how P_h , P_{DT} and P_{h+DT} were generated and compared during real hospital problems, the methodology for RQ2 also examined whether the professionals themselves learnt and how their understanding of problems developed over time. Human learning was assessed using several forms of observable evidence rather than relying on participants simply reporting that they had learnt. The questions generated by professionals (Q_h) were assessed for their relevance to the problem, ability to test assumptions and possible causes, attention to uncertainty, usefulness for identifying missing information, and consideration of stakeholder and value consequences. Their recorded reasoning provided external evidence of T_h , including critical and reflective thinking, logical reasoning, abstract reasoning and creative imagination. Judgement referred to the quality of the opinion formed from the available evidence, while decision quality concerned which option was selected and the reasons supporting that choice.

For reproducibility, the same structured case-assessment protocol was used at every formal assessment point. Professionals recorded, in a predefined sequence, their problem framing, questions generated, possible causes, evidence supporting or contradicting each cause, missing information, level of uncertainty, judgement, proposed decision and reasons, alternative solutions and value considerations. They also produced a causal or problem-relationship map showing how they believed the important factors were connected. Responses were completed before feedback was provided. A short semi-structured interview was used only where clarification of an existing response was necessary; it could not be used to introduce new reasoning after the assessment had ended. Standardised case assessments using previously unseen cases were conducted at baseline, at a predefined midpoint of the 6–8-month intervention, and after the principal field-learning period. The cases used at different time points were matched in advance for problem complexity, amount of available information, number of plausible competing causes, degree of uncertainty, operational constraints and value trade-offs. Matching criteria were fixed before participant assessment so that apparent improvement could not be created by giving easier cases later. Improvement on unfamiliar cases provided evidence that learning transferred beyond problems the professionals had already encountered.

Development from $M_t \rightarrow M_{t+1}$ was examined through repeated externalisations of how professionals understood a problem. Their representations included environmental conditions, building-system behaviour, occupancy and activities, causal relationships, human consequences, diagnostic evidence, potential solutions, stakeholder requirements and value implications. These were assessed for scientific correctness, completeness,

causal coherence, distinction between symptoms and causes, representation of uncertainty and integration of value relationships. A predefined anchored scoring rubric was used for each dimension. Three trained evaluators independently scored anonymised responses and mental-model externalisations without being told the participant's identity, whether the assessment was baseline, midpoint or post-intervention, or the expected direction of change. Inter-rater reliability was reported with appropriate agreement statistics and 95% confidence intervals, and scoring criteria were recalibrated if predefined reliability standards were not maintained. These externalisations were treated as observable evidence from which development of the internal mental model was inferred; the study did not claim to measure the mental model directly.

Intervention fidelity was also recorded so that a lack of improvement could be distinguished from failure to receive the intended intervention. Attendance at facilitated sessions, completion of episode-based reasoning records and reflections, and adherence to the prescribed learning sequence were documented. External professional training, substantial case discussions outside the study and other exposures that could influence learning were also recorded and considered in sensitivity analyses.

Measurement of Reciprocal Human–AI-DT Learning

The study also examined how professional learning contributed to subsequent AI-DT development and how AI-DT information contributed to further professional learning. For every real problem episode, a unique identifier linked the information available to the professional, questions generated, recorded reasoning, independent judgement, human–AI-DT interaction, final coupled judgement, field-verification evidence and subsequent reflection. All relevant events were timestamped so that the order of information, questioning, reasoning, judgement and learning could be reconstructed rather than inferred from participants' later recollection. Human-originated information supplied to the coupled AI-DT was retained with its provenance. Each item was labelled according to whether it arose from a professional question, observation, contextual information, causal hypothesis, stakeholder requirement or value consideration, and whether it was later verified, contradicted or remained unresolved. This prevented an unsupported human interpretation from becoming indistinguishable from objectively established machine information.

Ablation analysis was used offline to test whether particular human-derived information streams contributed to subsequent AI-DT performance. In simple terms, a completed AI-DT assessment was repeated after one predefined information class was deliberately removed, while the computational state and all remaining inputs were kept unchanged. For example, the analysis could remove professional-provided contextual information while leaving sensor, operational and other information unchanged. Performance with the complete information was then compared with performance after that information class had been removed. Consistent deterioration under the predefined comparison was interpreted as evidence that the removed information contributed to AI-DT performance under those

conditions. Ablation was performed only as an analytical procedure after the prospective episode; information was never withheld from live hospital problem solving for research purposes.

Mechanism analysis respected the temporal order of events. A high-quality question and a successful diagnosis occurring in the same episode were not considered sufficient proof that the question caused the improvement. The analysis examined whether the observed sequence was consistent with the proposed mechanism: information received → questioning → reasoning or information seeking → revised judgement or machine information → verified outcome → subsequent change in performance. This provided evidence consistent with the proposed reciprocal-learning mechanism without claiming direct proof of unobservable cognitive causation.

Measurement of Value-Oriented Problem Diagnosis and Solving

The same multidimensional P framework established in RQ1 was retained so that performance could be compared directly across the two research questions. It assessed problem framing, abnormality recognition, symptom-cause discrimination, spatial, temporal and operational reasoning, generation and discrimination of competing causes, identification of information required to reduce uncertainty, solution generation and consequence assessment, and value-oriented solution evaluation. For each problem episode, the independent evaluation panel assessed anonymised versions of P_h , P_{DT} and P_{h+DT} against the same independently adjudicated reference evidence. All outputs were converted into the same standard presentation format, with identifying wording, metadata and formatting removed where possible so that evaluators could not readily infer whether an output came from a professional, the AI-DT or the coupled system. The same 0–4 anchored rubric was used throughout, with periodic blinded recalibration to detect scoring drift.

The principal comparison was P_{h+DT} versus P_{DT} , which tested whether deliberate professional human learning and interaction added capability beyond AI-DT learning alone. The stronger complementarity test remained:

$$P_{h+DT} > \max(P_h, P_{DT})$$

Technical measures such as diagnostic accuracy, rank of the verified cause, false-positive rate, detection time and uncertainty calibration supplemented the multidimensional assessment. Repeated observations from the same professionals, teams, hospital locations and problem families were recognised as statistically related rather than independent and were handled using the multilevel statistical approach described in the subsequent analysis section. The principal outcome therefore remained improvement in value-oriented problem diagnosis and solving, not simply improved environmental prediction. Evidence for improvement required convergence across the structured performance scores, independently adjudicated real-world evidence, technical measures

and longitudinal learning records, while null findings and failures were retained rather than removed.

Evaluation of Reciprocal Human–AI-DT Learning, Complementarity and Capability Boundaries

The methodology for RQ2 examined whether deliberately coupling professional human learning with AI-DT learning produced additional capability beyond AI-DT learning alone, and under what conditions that benefit appeared, weakened or disappeared. The principal comparison was therefore between coupled professional–AI-DT performance P_{h+DT} and concurrent AI-DT-only performance P_{DT} for the same unresolved hospital problem. A stronger test of complementarity examined whether $P_{h+DT} > \max(P_h, P_{DT})$, meaning that the coupled system performed better than both the professional and the AI-DT acting independently. Complementarity could occur in some dimensions, such as causal diagnosis, identification of missing information or stakeholder-value reasoning, without necessarily occurring in every dimension.

Capability boundaries retained the problem, environmental, spatial, temporal and operational conditions established in RQ1 and added decision conditions relevant to human–AI interaction. These included diagnostic ambiguity, disagreement between professional and AI-DT judgement, machine uncertainty, novelty of the problem, incomplete information, time pressure and the importance of stakeholder or value considerations. A single success or failure was not treated as evidence of a boundary. Boundary claims required repeated and practically meaningful differences in performance under identifiable conditions, allowing the study to determine when reciprocal learning was useful and when it became unreliable or added little value.

The methodology targeted approximately 24 professionals and at least 80 independently adjudicable field problem episodes. Final sample adequacy was assessed through statistical simulation using empirical variation and dependence observed during RQ1 and the RQ2 pilot phase, targeting at least 90% statistical power at a two-sided $\alpha = 0.05$. The main statistical comparison was the within-episode difference between P_{h+DT} and P_{DT} , because both systems addressed the same real problem during the same period. Since the same professionals, teams, hospital locations and problem families could contribute several episodes, mixed-effects models were used to account for these related observations rather than treating every episode as completely independent.

Ordinal outcomes, such as the 0–4 performance scores, binary outcomes, such as whether a cause was correctly identified, and continuous outcomes, such as detection time, were analysed using corresponding mixed-effects models. The overall coupled-versus-AI-DT-only comparison was tested first, followed by individual performance dimensions using Holm adjustment to reduce the risk of false-positive findings from multiple testing. Results were reported using 95% confidence intervals, effect sizes and

exact P values. Statistical significance alone was insufficient; improvement also had to be practically meaningful and could not involve unacceptable deterioration in a safety-critical aspect of performance.

Mechanistic analyses examined whether changes in question quality, reasoning quality, judgement and externally represented mental-model quality helped explain improvements in coupled performance. Where appropriate, multilevel mediation analysis was used only for prespecified pathways in which the order of events had been established. Such findings were interpreted as evidence consistent with the proposed reciprocal-learning mechanism, rather than proof of internal cognitive processes that could not be observed directly. Sensitivity analyses examined uncertain reference diagnoses, missing data, measurement failure, incomplete exposure to the learning intervention and potential contamination from cross-team learning or external professional training. Failure analysis examined automation bias, inappropriate deference to AI-DT outputs, unjustified rejection of correct machine reasoning, overconfidence, confirmation bias and reciprocal reinforcement of incorrect causal explanations.

The null hypothesis was:

$$H_{02}: P_{h+DT} \leq P_{DT}$$

where P_{DT} represented performance under AI-DT learning without deliberately developed professional human learning and P_{h+DT} represented performance under deliberately coupled professional human and AI-DT learning. H_{02} therefore stated that deliberate professional learning and reciprocal human–AI-DT interaction did not produce meaningful improvement beyond AI-DT learning alone and that no reproducible additional benefit could be attributed to reciprocal development of the professional mental model and AI-DT computational representation.

The alternative hypothesis was:

$$H_{12}: P_{h+DT} > P_{DT}$$

under one or more identifiable conditions. Support for H_{12} required reproducible, practically meaningful and statistically supported improvement in one or more prespecified dimensions of value-oriented problem diagnosis and solving, together with evidence consistent with identifiable reciprocal human–AI-DT learning mechanisms. Reproducibility was supported through a version-controlled audit trail of human–AI interactions, AI-DT states, intervention materials, scoring instruments and statistical code. The AI-DT remained a decision-support system and could not autonomously alter hospital systems; all operational decisions remained with authorised personnel, and no harmful environmental condition was deliberately created.

The final scientific output was a human–AI-DT complementarity and boundary map showing how much deliberate professional learning added beyond AI-DT learning alone, which changes in questioning, reasoning, judgement and mental-model development were associated with that added capability, how verified human-derived information contributed to subsequent AI-DT learning, and under which conditions complementarity emerged or failed. RQ1 examined what progressive AI-DT learning, represented by $R_t \rightarrow R_{t+1} \rightarrow R_{t+2} \rightarrow \dots$, could achieve without deliberately developing professional human learning. The methodology for RQ2 then deliberately developed professional human learning and examined reciprocal development, in which human learning could generate new or revised I_m for subsequent machine learning, while AI-DT learning could generate new I_h for subsequent human learning. The reciprocal process was therefore represented as:

$(I_h, M_t) \rightarrow \text{human learning} \rightarrow M_{t+1} \rightarrow \text{new/revised } I_m \rightarrow (I_m, R_t) \rightarrow \text{machine learning} \rightarrow R_{t+1} \rightarrow \text{new } I_h \rightarrow (I_h, M_{t+1}) \rightarrow \text{further human learning} \rightarrow M_{t+2} \rightarrow \text{new/revised } I_m \rightarrow (I_m, R_{t+1}) \rightarrow \text{further machine learning} \rightarrow R_{t+2} \rightarrow \dots$

The methodology then determined whether this reciprocal developmental process produced greater value-oriented problem diagnosis and solving than progressive AI-DT learning alone, represented by:

$R_t \rightarrow R_{t+1} \rightarrow R_{t+2} \rightarrow \dots$

Ethical Considerations and Methodology Contribution to Knowledge

The methodology for RQ2 involved professionals' reasoning, learning, judgement and interaction with an AI-DT during real hospital work and therefore required ethical safeguards beyond conventional environmental monitoring. Participation was voluntary and subject to informed consent, with withdrawal permitted without employment consequences. Recruitment procedures minimised pressure arising from organisational hierarchy. Because the study generated potentially sensitive evidence about individual questioning, reasoning, judgement and decision-making, participant identities were pseudonymised, access to identifiable records was restricted, and individual research assessments were not used for staff appraisal, disciplinary action or employment evaluation. Interviews, reasoning records, interaction traces and mental-model externalisations were collected only to the extent necessary to answer the research question.

Particular attention was also given to patient, occupant, hospital-operational and cybersecurity information that could enter I_h or I_m . Personally identifiable information was minimised or removed where not scientifically necessary, while hospital building-system and operational information was handled under approved data-governance and access-control arrangements. Human-generated observations, interpretations and hypotheses remained distinguishable from verified evidence so that professional authority or confidence could not convert unsupported judgement into machine ground truth. The AI-DT

remained a decision-support research system and could not autonomously modify hospital systems. Operational decisions and corrective actions remained with authorised personnel through established hospital procedures. Clinical safety, infection-control requirements and urgent operational needs took precedence over research procedures, and no harmful indoor-environmental condition was deliberately created for research purposes.

The principal methodological contribution to knowledge was the development of an integrated prospective field framework for investigating reciprocal human–AI-DT learning over time, rather than simply comparing human and AI performance or observing their interaction at a single point. Its originality lay in combining prospective sequential evaluation with independent professional, AI-DT and coupled assessments of the same unresolved real problem before its cause was known; parallel AI-DT developmental lineages beginning from the same computational state; independent real-world verification and adjudication; longitudinal assessment of professional learning and mental-model development; provenance-tracked transfer of information between human and machine; and capability-boundary analysis within one continuing hospital field programme. The methodology therefore provides a reproducible way to determine whether, how, by how much and under what conditions deliberately developed professional human learning contributes to reciprocal AI-DT development and to greater value-oriented problem diagnosis and solving than progressive AI-DT learning alone. Its methodological contribution is consequently not merely a method for studying human–AI interaction, but a field approach for investigating reciprocal developmental complementarity, the mechanisms through which it emerges, and the conditions defining its capability boundaries.

Methods for Research Question 3:

Overview

The methodology for Research Question 3 continued the work carried out in RQ1 and RQ2 in the same large general hospital. RQ1 examined what the AI-DT could progressively learn when professional human learning was not deliberately developed. RQ2 then introduced deliberate professional learning and examined whether professionals and the AI-DT could learn from one another and achieve better value-oriented problem diagnosis and solving. RQ3 extended this established Professional Human ↔ AI-DT system by deliberately enabling hospital occupants to take part in the learning, diagnosis and solving of indoor-air and other indoor-environmental problems.

The main methodological change in RQ3 was therefore the introduction of occupant learning, participation and agency. Occupants were not treated simply as people who reported complaints or supplied information about how the environment felt. They were treated as participants who could describe what they experienced, receive understandable environmental information, ask questions, form judgements, explain what outcomes

mattered to them and learn from what was later established through physical investigation. The field infrastructure and procedures established in RQ1 and RQ2 were retained, including continuous indoor-environmental monitoring, problem-episode definitions, physical field investigation, independent reference diagnosis, test-before-learning procedures, computational version control and multidimensional evaluation of value-oriented problem diagnosis and solving.

The occupant-participatory intervention was introduced gradually across predefined occupant groups or hospital settings using a stepped-wedge design. In practical terms, different groups began participating at different planned times. Before a group received the intervention, eligible problems continued to be addressed through the established professional human–AI-DT system, represented by P_{h+DT} . After the intervention was introduced, occupants also contributed to the learning and problem-solving process, producing the occupant-participatory system represented by P_{o+h+DT} . For each eligible real hospital problem, assessments were made before the true cause was known and were later compared with evidence from physical field investigation and independent adjudication. Repeating this process across real problem episodes allowed the methodology for RQ3 to determine whether occupant participation added capability beyond the professional human–AI-DT system, how that additional capability emerged, whether occupants themselves developed greater capability and agency, and under which problem, occupant, environmental, spatial, temporal and operational conditions occupant-enabled complementarity strengthened, weakened or failed to emerge.

Occupant Human-Learning Framework

RQ3 examined not only whether information provided by occupants improved problem diagnosis and solving, but also whether occupants themselves developed greater capability to participate meaningfully in that process. An explicit occupant human-learning framework was therefore required to distinguish the occupant as a learning participant from the occupant merely as a source of complaints, perceptions or other information. The methodology did not introduce a different theory of human learning for occupants. Instead, it applied the human-learning framework established in RQ2 to the occupant, using the subscript o to identify the occupant-learning pathway. Occupants were themselves human participants; the subscript o was therefore not intended to distinguish occupant learning from human learning. Rather, it distinguished the occupant human in RQ3 from the professional human in RQ2, where the human-learning framework referred specifically to hospital professionals whose roles involved managing, diagnosing and solving indoor-air and other indoor-environmental problems.

The occupant human-learning process was represented as:

$$Q_{h,o} = f(I_{h,o}, M_{t,o})$$

$$T_{h,o} = f(Q_{h,o})$$

$$L_{h,o} = f(T_{h,o})$$

$$\text{Exp}_{h,o} = I_{h,o} \times L_{h,o}$$

$$M_{t+1,o} = f(M_{t,o}, \text{Exp}_{h,o})$$

where $I_{h,o}$ represented information received by the occupant from any source, including personal observations and perceptions, other occupants, professionals, environmental information and the AI-DT; $M_{t,o}$ represented the occupant's existing mental model at time t ; $Q_{h,o}$ represented the questions generated from that information and existing understanding; $T_{h,o}$ represented the occupant's thinking process; $L_{h,o}$ represented human learning; $\text{Exp}_{h,o}$ represented the knowledge, understanding and skills developed through the interaction between information and learning; and $M_{t+1,o}$ represented the updated occupant mental model.

The complete occupant human-learning process was therefore:

$$(I_{h,o}, M_{t,o}) \rightarrow Q_{h,o} \rightarrow T_{h,o} \rightarrow L_{h,o} \rightarrow \text{Exp}_{h,o} \rightarrow M_{t+1,o}$$

As in RQ2, these were process equations rather than statistical regression equations. The multiplication sign in $\text{Exp}_{h,o} = I_{h,o} \times L_{h,o}$ represented interaction rather than numerical multiplication. Information received by the occupant was also not automatically treated as knowledge or understanding. The occupant still had to question, think about, interpret and learn from that information before it could contribute to $\text{Exp}_{h,o}$ and potentially refine the existing mental model from $M_{t,o}$ to $M_{t+1,o}$. This distinction was important because an occupant's immediate experience of an indoor environment formed part of the information entering the learning process, whereas $\text{Exp}_{h,o}$ represented what the occupant subsequently learnt from processing information.

Internal thinking and mental models could not be observed directly. The methodology therefore examined their external evidence. $Q_{h,o}$ was captured through the questions occupants generated. Evidence of $T_{h,o}$ came from how occupants interpreted information, distinguished observations from assumptions, considered possible explanations and provided reasons for their judgements. $L_{h,o}$ was indicated by defensible changes in knowledge, understanding, judgement or capability to participate following engagement with information and verified outcomes. Development from $M_{t,o} \rightarrow M_{t+1,o}$ was therefore inferred from repeated changes in how occupants externally represented their understanding of a problem, including what they experienced, what they thought might explain it, what information they considered important and what outcomes mattered to them.

Importantly, an occupant's situated observation or perception was not automatically treated as an established environmental fact or verified cause. For example, "the air smells musty" was retained as a legitimate perceived-air-quality observation and therefore part of the

information available for problem diagnosis, whereas “mould is growing in the ventilation system” was treated as a causal hypothesis requiring measurement, inspection, field investigation or independent adjudication. This distinction allowed occupants’ situated experiences to contribute meaningfully to problem diagnosis and solving without converting subjective perception or unverified interpretation into ground truth.

Integrated Occupant–Professional–AI-DT Learning Architecture

Building on the occupant human-learning framework above, RQ3 examined how what occupants learnt could contribute to the wider problem-solving system and how information generated by that system could, in turn, support further occupant learning. RQ3 therefore extended the reciprocal architecture established in RQ2: Professional Human–AI-DT → Occupant–Professional Human–AI-DT. The integrated architecture also allowed direct information exchange between the occupant and AI-DT where appropriate. The occupant was therefore not required to communicate with the AI-DT only through the professional; occupant contributions could become information for the professional, the AI-DT, or both.

The purpose was not simply to increase the amount of information available. Occupants could contribute situated experiences, contextual relationships, questions, stakeholder requirements, sacrifices and consequences that might be unavailable, incompletely represented or inadequately interpreted by environmental sensors, professionals or the AI-DT. Their contribution could therefore add a perspective arising from actually experiencing the indoor environment and the consequences of the problem.

When an occupant’s mental model developed from $M_{t,o}$ to $M_{t+1,o}$, that development could change what the occupant noticed, questioned, interpreted or communicated. An occupant contribution could therefore become new or revised I_h for the professional, new or revised I_m for the AI-DT, or both. The professional and AI-DT then processed the information through the human- and machine-learning pathways already established in RQ2. The integrated reciprocal pathway was represented as:

$(I_{h,o}, M_{t,o}) \rightarrow \text{occupant learning} \rightarrow M_{t+1,o} \rightarrow \text{occupant contribution} \rightarrow \{\text{new/revised } I_h \rightarrow (I_h, M_t) \rightarrow \text{human learning} \rightarrow M_{t+1}; \text{new/revised } I_m \rightarrow (I_m, R_t) \rightarrow \text{machine learning} \rightarrow R_{t+1}; \text{or both pathways}\} \rightarrow \text{new } I_{h,o} \rightarrow \text{further occupant learning} \rightarrow M_{t+2,o} \rightarrow \dots$

This expression represented a reciprocal and iterative information-and-learning pathway rather than a compulsory linear sequence. It did not imply that occupant, professional and AI-DT learning always occurred in this order, simultaneously or at the same rate. Instead, learning by one part of the system could generate new or revised information that subsequently contributed to learning by another. New $I_{h,o}$ could similarly arise from the professional, the AI-DT or both.

Occupant-derived information was not automatically treated as verified fact; observations, interpretations and hypotheses retained their provenance and verification status. Direct

entry of occupant-derived information into I_m therefore did not convert it into machine ground truth. Its status as observation, interpretation, hypothesis or subsequently verified information remained identifiable throughout the learning process. This allowed occupant contributions to influence professional and AI-DT reasoning without giving an unsupported interpretation the status of established evidence.

The phenomenon investigated was consequently participatory complementarity rather than crowdsourced sensing. The methodological question was not merely whether occupants could provide additional data, but whether deliberately developing their capability to question, interpret, judge and learn enabled them to contribute meaningfully to value-oriented problem diagnosis and solving, and whether those contributions improved the established professional human–AI-DT system.

RQ3 commenced from the professional human–AI-DT system established and prospectively evaluated in RQ2. The environmental-monitoring network, problem definitions, field-verification procedures, professional human-learning framework, independent adjudication process and value-oriented performance measures were retained. This continuity was important because differences observed during RQ3 could then be linked more convincingly to the introduction of deliberate occupant learning, participation and agency rather than to changes in the underlying monitoring, diagnostic or evaluation procedures.

Two AI-DT lineages were maintained from the same terminal computational state established in RQ2. One continued within the professional human–AI-DT system without deliberately developed occupant-learning input, while the other could additionally receive appropriately identified and, where required, verified information, contextual relationships and stakeholder-value requirements arising from deliberate occupant participation. Both therefore began RQ3 with the same prior machine development, while the professional-learning foundation established in RQ2 was retained. Together with the stepped-wedge introduction of the occupant intervention, this provided a structured comparison between the established P_{h+DT} system and the occupant-participatory P_{o+h+DT} system.

Routine complaints and ordinary communication from occupants were not suppressed in the comparison condition, because doing so would be artificial and potentially unethical in a functioning hospital. The intervention instead concerned the additional contribution created by deliberately developing occupant learning, structured questioning, interpretation, judgement, feedback and agency beyond what occupants would ordinarily contribute through routine reporting.

Field Experimental Design

RQ3 used a cluster-randomised stepped-wedge field design because the intervention aimed to develop occupants' learning, participation and agency, capabilities that could not realistically be withdrawn once developed. The intervention was therefore introduced

progressively across approximately eight hospital clusters, defined as physically and operationally coherent areas in which occupants, indoor-environmental conditions and problem-management processes were sufficiently connected for meaningful intervention and evaluation. Cluster boundaries and eligibility criteria were prespecified before randomisation.

The field programme lasted approximately eight months. A four-week pilot and commissioning period, excluded from the primary effectiveness analysis, tested intervention delivery, participation procedures and data capture. This was followed by five six-week stepped-wedge periods, providing approximately 30 weeks of prospective evaluation. During the first formal period, all clusters operated under the professional human–AI-DT system established in RQ2. At the end of each period, approximately two clusters crossed to the occupant-participatory condition according to a computer-generated randomisation schedule fixed before prospective evaluation, until all clusters had received the intervention.

The stepped-wedge design controlled when structured occupant learning and participation were introduced, not when indoor-environmental problems occurred. All episodes were naturally occurring real hospital problems arising during normal operation. Pollutant releases, microbial contamination, thermal discomfort, noise, lighting problems, ventilation faults and occupant complaints were not created, scheduled, induced or manipulated. Thus, the experimental element concerned intervention timing rather than environmental-problem occurrence.

Naturally occurring problems could nevertheless introduce bias because problem type, frequency, severity and complexity could differ across cluster-periods. The intervention could also increase occupant awareness and reporting. The methodology therefore did not assume that pre- and post-intervention episodes were automatically comparable. The same prospectively defined episode-detection and eligibility criteria were applied throughout; all candidate, excluded and unresolved episodes were logged with reasons; and routine monitoring, existing hospital reporting pathways and other detection mechanisms remained unchanged. Calendar period was included in analysis to distinguish intervention effects from season, hospital operations, continuing professional experience and ongoing AI-DT development.

Potential cross-cluster contamination was also addressed. Cluster selection aimed to minimise unnecessary movement of recurrent occupants and staff between conditions, while professional crossover and other spillover were documented for sensitivity analyses. Episode characteristics, including problem family, severity, environmental condition, location, timing, operational context, initially available information and diagnostic complexity, were compared across conditions. Meaningful imbalances were incorporated into multilevel analyses, while prespecified stratified and sensitivity analyses examined comparable problem types. An additional sensitivity analysis was restricted to episodes

detected through routine monitoring or pre-existing reporting pathways independent of the intervention.

A within-episode comparison was embedded within the stepped-wedge design. For each eligible intervention-period episode, the established professional human–AI-DT system first produced a locked P_{h+DT} assessment using information ordinarily available while the problem remained unresolved. Structured occupant contributions were then introduced, followed by a second locked P_{o+h+DT} assessment before the actual cause was known. During comparison periods, an equivalent time-matched second review occurred without structured occupant participation. This helped distinguish improvement attributable to occupant participation from improvement caused simply by reconsidering the problem.

Occupant Participants: Occupants were defined as people who directly experienced the hospital environment but were not professionally responsible for managing its indoor-environmental problems. Eligible participants included adult healthcare workers who routinely occupied participating areas, adult patients clinically able to participate, and carers or visitors with repeated exposure to relevant environments.

Longitudinal analysis of occupant learning focused primarily on recurrent occupants, because development from $M_{t,o} \rightarrow M_{t+1,o}$ required repeated opportunities to receive information, question, think, learn and apply previous experience. Approximately 80 recurrent occupants were targeted across participating clusters. A planning target of approximately 96 independently adjudicable problem episodes was used for feasibility, while final sample adequacy was determined through simulation incorporating cluster structure, repeated occupants, repeated hospital locations, expected problem frequency and preliminary variance and dependence estimates from RQ1, RQ2 and the RQ3 pilot. The number of occupants alone was therefore not treated as the effective statistical sample size.

Patients participated only where they were adults able to provide informed consent and participation would not interfere with clinical care. Clinically unstable patients and people unable to provide informed consent were excluded from longitudinal learning assessments. Transient occupants could provide relevant information about a particular episode, but a single encounter was not used as evidence of progressive occupant learning.

Occupant Learning and Participation Intervention: The intervention was designed to develop capability and agency, rather than simply increase the number of complaints submitted. Occupants received a short structured orientation distinguishing direct observation or experience from interpretation and causal hypothesis. They were supported to describe what they experienced, where and when it occurred, what they were doing at the time, how the condition affected them, what questions arose and what outcome would be useful.

Occupants were not expected to become substitute environmental engineers or diagnose technical faults beyond their competence. Their role was to contribute situated knowledge, generate meaningful questions, interpret information appropriately and articulate stakeholder-value requirements. A secure participation interface therefore distinguished among situated experience, such as odour, stuffiness, draught, thermal discomfort, noise or lighting discomfort; interpretation or hypothesis, meaning what the occupant thought might explain the condition; and value judgement, meaning what outcome mattered, what sacrifice or inconvenience was unacceptable and what competing requirement should be considered.

Information returned by the professional–AI-DT system was communicated in accessible language and excluded sensitive clinical, cybersecurity or building-security information. Occupants could question that information rather than being expected merely to accept it. Intervention exposure was documented, including completion of the orientation, use of the structured participation process and receipt of verified feedback, so that failure of the intervention could be distinguished from failure to receive it as intended.

Integration With Environmental Monitoring: RQ3 retained the environmental-monitoring platform established through RQ1 and RQ2. Available objective information could include chemical, particulate and biological indoor-air measurements; temperature and humidity; airflow and pressure relationships; acoustic and lighting conditions; occupancy and space utilisation; HVAC operation; cleaning and maintenance activities; and relevant outdoor conditions.

Perceived indoor-air quality remained complementary to these objective measurements. Reports of air acceptability, freshness or stuffiness, odour intensity and odour character were linked to where and when the experience occurred and to the activity taking place. RQ3 added structured occupant interpretation of this context. For example, an occupant might identify that a draught occurred only at one bed location or that an odour repeatedly appeared during a particular cleaning activity. Such information could generate testable questions and guide targeted investigation, but it did not replace measurement, field investigation or independent diagnosis.

Prospective Problem-Episode Procedure: Eligible problem episodes followed the definitions established in RQ1 and RQ2 and were registered before their underlying causes were known. Episodes entered the study only when naturally encountered through normal hospital operation and when they satisfied the predefined eligibility criteria. Researchers neither delayed a necessary response nor allowed a problem to continue merely to obtain research data.

To minimise ascertainment and selection bias arising from natural problem occurrence, candidate episodes were recorded prospectively regardless of whether they occurred in a comparison or intervention period and regardless of whether the initial assessment

suggested that occupant participation would be useful. Eligibility could not be changed after the professional or AI-DT assessment was seen. Episodes with unsuccessful diagnoses, no apparent benefit from occupant participation, or unresolved reference diagnoses remained in the study record rather than being selectively discarded.

During comparison periods, the established professional human–AI-DT system generated P_{h+DT} using the RQ2 procedure. Routine complaints and ordinary occupant communication remained available through normal hospital processes, but deliberate occupant learning and structured participation were not introduced.

During intervention periods, the professional human–AI-DT system first produced and locked an episode-specific P_{h+DT} assessment. Directly affected occupants then contributed situated experiences, questions, interpretations and value requirements through the structured participation process. The professional and AI-DT could interrogate those contributions, request clarification and identify additional information or investigation where justified. The occupant-participatory system then produced the second locked assessment, P_{o+h+DT} .

For each prespecified performance dimension, the within-episode difference was represented as:

$$\Delta P_o = P_{o+h+DT} - P_{h+DT}$$

The equivalent time-matched repeat-review change during comparison-period episodes was used to estimate improvement attributable simply to reconsidering the problem. ΔP_o was therefore not interpreted in isolation as proof that the current occupant contribution caused the observed difference. Intervention effects were determined from the stepped-wedge, within-episode and longitudinal evidence together. Interpretation also considered whether intervention and comparison periods contained comparable problem mixes. Where they did not, conclusions were based on adjusted, stratified and sensitivity analyses rather than attributing raw differences directly to the occupant intervention.

Field Verification and Reference Diagnosis: After prospective assessments had been locked, each episode underwent the same physical field-investigation and independent reference-diagnosis procedures established in RQ1 and RQ2. Depending on the problem, investigation could include portable pollutant measurements, spatial mapping, airflow and pressure testing, equipment inspection, chemical or biological sampling, thermal, acoustic or lighting measurements, moisture investigation, and review of activities, maintenance or adjoining spaces.

Occupant information could generate a hypothesis but could not verify itself. A report of a musty odour, for example, could justify investigation of moisture or microbial sources, but the technical cause required independent evidence. Where an authorised corrective action occurred, the established verification sequence remained:

Measured problem → competing explanations → targeted investigation → identified cause
→ authorised corrective action → predicted response → observed response

The independent reference-diagnosis panel subsequently classified causes as verified, probable or unresolved using the same evidential principles employed in RQ1 and RQ2. The reference process remained independent of whether improved performance was expected from occupant participation, and unresolved causes were retained as unresolved rather than forced into a definitive diagnosis.

Technical diagnosis and stakeholder-value evidence were kept conceptually distinct. Physical, environmental and operational evidence established what was happening and why, whereas occupant and other stakeholder evidence established what outcomes, sacrifices or unintended consequences mattered. This prevented occupant participation from becoming its own technical reference standard.

Feedback, Learning and Mental-Model Development: After independent investigation and adjudication, participating occupants received concise feedback explaining what had been established, what remained uncertain, what action had been taken where appropriate and how their contribution related to the investigation. They then completed a brief structured reflection on what they had interpreted appropriately, which assumptions were unsupported, what information proved important and what they would consider differently in a future problem.

This process provided new $I_{h,o}$ that could support further questioning, thinking and learning. Consistent with the occupant human-learning framework:

$$Exp_{h,o} = I_{h,o} \times L_{h,o}$$

and

$$M_{t+1,o} = f(M_{t,o}, Exp_{h,o})$$

Repeated changes in observable questioning, interpretation, judgement and problem representation were therefore used as evidence of developing occupant capability. Where occupant participation revealed a verified contextual relationship or stakeholder requirement not previously represented, its provenance was retained and it could contribute to subsequent professional and AI-DT learning. Unsupported occupant hypotheses remained labelled as hypotheses rather than becoming verified machine information.

The use of naturally occurring problems therefore strengthened ecological validity because the methodology examined the system under genuine hospital conditions, but it did not provide the balance of problem types that would be expected from experimentally staged cases. The methodology addressed this limitation through randomised intervention timing, prospective and condition-independent episode registration, fixed eligibility criteria,

preservation of all eligible successes and failures, characterisation and statistical adjustment of problem mix, and sensitivity analyses restricted to comparable or independently detectable episodes. Consequently, claims of occupant-enabled complementarity were made only where the observed advantage remained reproducible after these potential natural-occurrence biases had been considered.

Measurement of Occupant Capability and Agency

RQ3 distinguished occupant empowerment from simply being consulted or feeling satisfied with participation. Empowerment was operationalised as the combination of capability and agency. Capability referred to what occupants progressively became able to contribute to problem diagnosis and solving, whereas agency referred to whether the hospital problem-solving system gave those contributions a genuine opportunity to influence investigation, judgement or action. Both were required because an occupant might develop useful understanding but have no opportunity to influence the process, or might be invited to participate without having developed the capability to contribute meaningfully. The conceptual learning pathway through which occupant capability could develop was defined earlier in the *Occupant Human-Learning Framework* and was not repeated here.

Occupant capability was assessed longitudinally using evidence from naturally occurring problem episodes and standardised unfamiliar transfer cases. The transfer cases provided comparable problems at predefined assessment points, while the real episodes showed whether learning could be applied during actual hospital practice. Assessment examined the clarity and contextual usefulness of situated-experience descriptions, relevance of questions, ability to distinguish observation from causal hypothesis, interpretation of available information, quality of judgement, articulation of useful outcomes and competing sacrifices, and evidence of learning from independently verified outcomes. Predefined anchored criteria were used so that development was judged consistently rather than through unrestricted subjective impressions. Repeated changes in these observable externalisations provided evidence of development from $M_{t,o} \rightarrow M_{t+1,o}$; the internal mental model itself was not claimed to have been directly measured. The procedures for participant selection, repeated exposure, feedback and reflection that supported this longitudinal assessment were specified in the *Field Experimental Design and Feedback, Learning and Mental-Model Development* sections.

Agency was assessed primarily from the prospective audit trail rather than from occupants merely reporting that they felt heard. The study examined whether an occupant had an opportunity to contribute and whether that contribution subsequently entered professional or AI-DT reasoning, generated a new measurement or investigation, changed problem framing or the ranking of possible causes, influenced solution evaluation, introduced a previously overlooked value consideration, or otherwise affected the problem-solving pathway. Receipt of feedback was also recorded because meaningful participation required occupants to learn what subsequently happened. Self-reported confidence,

perceived influence and satisfaction were retained as secondary information but did not substitute for demonstrated agency. How occupant contributions entered the professional and AI-DT learning processes, while retaining their provenance and verification status, was established in the *Integrated Occupant–Professional–AI-DT Learning Architecture and Continuity from RQ1 and RQ2* section.

System-Level Value-Oriented Problem Diagnosis and Solving

The multidimensional P framework established in RQ1 and retained in RQ2 was used to compare P_{h+DT} with P_{o+h+DT} . This preserved comparability across the three research questions while allowing RQ3 to determine whether structured occupant participation added capability to the already established professional human–AI-DT system. Standardised, anonymised outputs were evaluated by the blinded evaluation panel using the established 0–4 anchored rubric, covering problem framing, symptom-cause discrimination, competing explanations, spatial, temporal and operational reasoning, identification of missing information, solution generation and consequence assessment, and value-oriented judgement. Technical measures such as diagnostic accuracy, ranking of the verified cause, false-positive rate and uncertainty calibration remained supplementary rather than replacing the broader assessment of problem-solving quality. The generation and locking of P_{h+DT} and P_{o+h+DT} , the time-matched repeat review, and the independent field verification against which they were judged were described in the *Prospective Problem-Episode Procedure* and *Field Verification and Reference Diagnosis* sections.

Value orientation was particularly important in RQ3 because occupants could reveal consequences that technical measurements alone could not establish. A technically effective solution could still provide poor value if it created unacceptable disruption to sleep, comfort, privacy, clinical work or another legitimate stakeholder requirement. Conversely, occupant preference could not override non-negotiable requirements such as safety, infection control or essential clinical operation. Evaluation therefore examined how technical effectiveness and stakeholder value were reconciled rather than automatically privileging either.

Mechanisms and Capability Boundaries

Mechanism analysis examined how occupant participation contributed to any observed improvement, rather than stopping at whether performance changed. Contributions were timestamped and provenance-labelled so that the sequence from occupant observation or question, through new information or revised reasoning, to the final judgement could be reconstructed. Plausible mechanisms included previously unavailable contextual information, recognition of spatial or temporal patterns, diagnostically useful questions, identification of disagreement between perception and measurement, overlooked stakeholder requirements, unintended consequences, or stimulation of professional or AI-DT reconsideration.

Temporal order was required: an occupant contribution and a correct diagnosis occurring in the same episode were insufficient unless the contribution plausibly preceded and informed the relevant reasoning or investigation. Where appropriate, multilevel mediation analysis provided additional evidence consistent with these pathways without claiming direct proof of unobservable cognitive causation. The safeguards against attributing improvement merely to repeated consideration of a problem, and against bias arising from naturally occurring problem episodes, were specified in the *Field Experimental Design* and *Prospective Problem-Episode Procedure* sections and were applied when interpreting these mechanisms.

Boundary analysis extended the conditions examined in RQ1 and RQ2 to include occupant and participatory conditions, such as occupant role, familiarity with the environment, previous participation, perceptibility of the problem, communication capability, disagreement among occupants, disagreement with professional judgement and conflict between experiential and sensor information. A capability boundary required recurrent, practically meaningful patterns across multiple episodes rather than a single success or failure. This allowed RQ3 to establish not only whether occupant participation helped, but when, how and for whom it added value, and when it added little or could potentially mislead the wider system.

Statistical Analysis, Sensitivity, and Hypothesis Evaluation

Approximately 80 occupants who repeatedly spent time in participating hospital areas, across eight clusters, and at least 96 naturally occurring hospital problem episodes were targeted. Repeated participation allowed changes in occupant learning and capability to be examined over time. Each problem episode required sufficient independent field evidence for the reference-diagnosis panel to classify its underlying cause as verified, probable or unresolved. Final sample adequacy was assessed through simulation using episode frequencies, variance components and dependence observed in RQ1, RQ2 and the RQ3 pilot. The simulation accounted for repeated episodes within clusters, recurring occupants and professionals, and variation among problem families rather than assuming that all observations were statistically independent.

The primary system-level comparison examined differences in value-oriented problem diagnosis and solving between the occupant-participatory system (P_{o+h+DT}) and the established professional human–AI-DT system (P_{h+DT}). Mixed-effects models accounted for related observations involving the same clusters, problem families, professionals or occupants. Intervention condition and calendar period were included as fixed effects, while appropriate random effects represented repeated and clustered observations. Bias arising from naturally occurring problems, including differences in problem type, frequency, severity and reporting, was addressed using the procedures established in the Field Experimental Design.

The embedded within-episode analysis compared the change from the first locked P_{h+DT} assessment to the subsequent P_{o+h+DT} assessment with the time-matched change produced by reviewing a problem twice during comparison periods without structured occupant participation. This difference-in-differences analysis helped distinguish improvement attributable to occupant participation from improvement caused merely by reconsidering the same problem. Findings were interpreted alongside the cluster-randomised stepped-wedge analysis rather than as independent proof of causation.

Ordinal outcomes, including the established 0–4 performance scores, were analysed using ordinal mixed-effects models; binary outcomes using generalised linear mixed-effects models; and continuous outcomes using linear mixed-effects models where appropriate. Longitudinal development in occupant capability and agency was examined using mixed-effects growth models based on repeated real problem episodes and standardised transfer cases. Results were reported with 95% confidence intervals, effect sizes and exact P values. The global system-level comparison preceded confirmatory dimension-specific analyses, with Holm adjustment controlling family-wise error from multiple comparisons. Statistical significance alone was insufficient; improvement also had to be practically meaningful without unacceptable deterioration in safety, infection control or essential clinical function.

Sensitivity analyses examined incomplete participation, missing environmental information, uncertain reference diagnoses, unequal episode frequencies, differences among healthcare workers, patients and carers, variation in the naturally occurring problem mix and possible cross-cluster contamination. Professional movement, cross-cluster discussion and exposure to occupant-derived information were documented because learning could transfer between hospital areas. Failure analysis retained episodes in which occupant participation produced no benefit or worsened performance, including misleading occupant interpretations, excessive professional deference to subjective experience, dismissal of valid occupant information, conflicting occupant experiences, overreaction to technically unimportant concerns and reciprocal reinforcement of incorrect explanations. These failures contributed directly to identifying the boundary conditions of participatory complementarity.

Reproducibility was supported through the computational and field audit trail established in RQ1 and RQ2. Pseudonymised participant records, timestamps, environmental information, occupant and professional contributions, AI-DT states, locked assessments and verification outcomes were prospectively linked. Interface versions, learning materials, scoring instruments, statistical code and protocol amendments were version controlled. The stepped-wedge allocation sequence, assessment rubrics, mechanism definitions, boundary conditions and statistical analysis plan were finalised before locked analysis.

The primary system-level null hypothesis was:

$$H_{03}: P_{o+h+DT} \leq P_{h+DT}$$

The alternative hypothesis was:

$$H_{13}: P_{o+h+DT} > P_{h+DT}$$

Here, P_{h+DT} represented the value-oriented problem-diagnosis-and-solving performance of the established professional human–AI-DT system, while P_{o+h+DT} represented performance after structured occupant participation was added. The alternative applied under one or more identifiable conditions. Occupant capability and agency were prespecified longitudinal outcomes rather than mathematical components of the performance inequality. Support for the wider RQ3 proposition therefore required reproducible, statistically supported and practically meaningful improvement in at least one prespecified performance dimension, together with evidence of increased occupant capability and/or demonstrated agency and an identifiable pathway through which occupant participation contributed to the observed change.

The AI-DT remained a decision-support system without autonomous authority to modify ventilation, pressure, filtration, temperature, lighting or other hospital systems. Occupant suggestions could not directly alter operational systems, and authorised hospital personnel retained decision-making responsibility. Participation was voluntary, patient care took precedence, and no harmful environmental condition was deliberately created for research purposes. The safety and non-staging provisions established in the Field Experimental Design remained applicable throughout RQ3.

The final scientific output was an occupant empowerment, complementarity and boundary map identifying what occupants contributed beyond information already available to professionals and the AI-DT, whether occupants became more capable and agentic participants through learning, how their contributions influenced professional or machine reasoning, how stakeholder usefulness and sacrifices affected solution evaluation, and under what conditions participation added value, added little value or potentially weakened value-oriented problem diagnosis and solving.

The three research questions formed a cumulative methodological progression. RQ1 investigated progressive AI-DT learning through $R_t \rightarrow R_{t+1} \rightarrow R_{t+2} \rightarrow \dots$. RQ2 introduced deliberate professional human learning and reciprocal professional human–AI-DT learning, whereby changes in M_t could generate new or revised I_m for subsequent machine learning, while AI-DT outputs could become new I_h for further professional learning. RQ3 extended this architecture by introducing deliberate occupant learning, participation and agency:

$(I_{h,o}, M_{t,o}) \rightarrow \text{occupant learning} \rightarrow M_{t+1,o} \rightarrow \text{occupant contribution} \rightarrow \{\text{new/revised } I_h \rightarrow (I_h, M_t) \rightarrow \text{human learning} \rightarrow M_{t+1}; \text{ new/revised } I_m \rightarrow (I_m, R_t) \rightarrow \text{machine learning} \rightarrow R_{t+1}; \text{ or both pathways}\} \rightarrow \text{new } I_{h,o} \rightarrow \text{further occupant learning} \rightarrow M_{t+2,o} \rightarrow \dots$

The methodological progression was therefore from machine learning, through professional human–AI-DT reciprocal learning, to occupant-participatory reciprocal learning and agency. The RQ3 expression did not imply that occupant, professional and AI-DT learning occurred simultaneously, in a compulsory sequence or at the same rate. Rather, it represented how learning by one part of the system could generate new or revised information that subsequently contributed to learning elsewhere in the system.

Ethical Considerations and Methodology Contribution to Knowledge

RQ3 was designed to protect occupants, professionals and patients while examining whether deliberately developed occupant learning, participation and agency could improve value-oriented diagnosis and solving of indoor-air and other indoor-environmental problems. Participation was voluntary, with informed consent obtained where required, and withdrawal did not affect employment, access to hospital services or clinical care. Patient participation remained subordinate to clinical priorities, and no research activity was permitted to interfere with treatment. Participant information was pseudonymised, access controlled, and professional or occupant contributions were not used for employment appraisal. No harmful environmental condition was deliberately created, prolonged or manipulated for research. Only naturally occurring problem episodes were investigated. The AI-DT remained a decision-support system without autonomous authority to modify ventilation, pressure, filtration, temperature, lighting or other hospital systems. Occupant suggestions could not directly trigger operational changes. Decisions remained subject to authorised hospital governance, infection-control requirements, established safety procedures and appropriate professional judgement.

The principal methodological contribution of RQ3 was a prospective field framework for investigating occupant-participatory reciprocal learning and agency within an already established professional human–AI-DT system. Rather than treating occupants merely as sources of complaints, perceptions or crowdsourced sensor information, the methodology deliberately developed and longitudinally evaluated their capability to question, interpret, judge, learn and participate in problem diagnosis and solving. The stepped-wedge design enabled structured occupant participation to be introduced progressively across hospital clusters while preserving comparison with the established P_{h+DT} system. Locked assessments, independent physical verification, reference diagnosis, blinded performance evaluation and embedded within-episode comparisons strengthened attribution of observed changes to occupant participation rather than simple reconsideration or unverified claims.

A further contribution was the explicit tracing of how occupant learning generated new or revised I_h , I_m , or both, while preserving the provenance and verification status of observations, interpretations and hypotheses. By integrating longitudinal assessment of occupant capability and agency with system-level performance, RQ3 could determine whether occupants became more capable participants, what they contributed beyond

professionals and the AI-DT, how those contributions influenced human or machine reasoning, when participation improved value-oriented problem solving, and where the boundaries of participatory complementarity emerged.

..... Chapter 4

Research Findings

Findings for Research Question 1:

Overview

The 12-month prospective study showed that the progressively learning AI-based digital twin (AI-DT) developed greater capability for value-oriented diagnosis and solving of indoor-air and other indoor-environmental problems than the frozen non-learning digital-twin baseline. Both systems began from the same initial computational representation, R_0 , received the same information available at the time of each unresolved problem, and produced their assessments before the actual cause was known. Only the learning AI-DT was subsequently allowed to learn from independently verified problem outcomes. This made it possible to determine whether better performance resulted from progressive learning rather than from having more information, a better starting point or knowledge of what happened afterwards.

During the study, 139 possible problem episodes were identified. Of these, 121 met the predefined requirements for inclusion in the study. For 105 of these 121 problem episodes, the investigation produced strong enough evidence to establish what had caused the problem. For another 10 episodes, the evidence pointed to a likely cause but was not strong enough to confirm it conclusively. For the remaining 6 episodes, the available evidence was insufficient to determine the cause with confidence. The problems covered six broad types, namely air pollutants, odour or chemical problems, moisture or biological problems, ventilation or pressure problems, thermal discomfort, and acoustic or lighting problems. They also occurred in different locations, at different times and under different hospital operating conditions.

Overall, the learning AI-DT achieved an average performance score of 3.17 out of 4.00, compared with 2.54 for the frozen non-learning digital twin, which started with the same initial information but was not allowed to learn from previous problem episodes. The adjusted improvement was 0.63 points (95% CI 0.49–0.77, $P < 0.001$), representing a substantial difference in performance (effect size 0.82). A statistical analysis that accounted for similarities among problem episodes confirmed that the learning AI-DT was substantially more likely to achieve a higher performance rating (OR 2.91, 95% CI 2.22–3.82). No safety-critical performance area became worse. For the 105 problem episodes in which the investigation established the cause with strong evidence, the learning AI-DT identified the correct cause as its first choice in 77.1% of the cases, compared with 56.2%

for the frozen non-learning digital twin. Improvement was greatest for recurring and clearly interpretable problems and smaller for complex, intermittent or multi-causal problems. The results therefore rejected H_{01} and supported H_{11} , while also showing where the AI-DT's capability remained limited.

Prospective Problem Episodes and Reference Diagnosis

A total of 121 indoor-air and other indoor-environmental problem episodes were included in the final analysis. These episodes represented situations in which the actual indoor environmental condition differed meaningfully from the condition required to support hospital activities or occupants. Across these episodes, the progressively learning AI-based digital twin (AI-DT) consistently outperformed the frozen non-learning digital twin in identifying the underlying cause and in ranking plausible explanations. The 121 episodes comprised 26 particulate or gaseous indoor-air pollutant problems, 17 VOC or odour problems, 14 moisture or biological problems, 22 ventilation or pressure-control problems, 23 thermal problems, and 19 acoustic or lighting problems. Performance differences between the two systems were therefore observed across a broad range of indoor-environmental problems rather than within a single problem type.

Of the 139 potential problem episodes initially registered, 121 were retained for analysis. The remaining 18 did not satisfy the predefined eligibility requirements: seven were continuations of already registered problems, six did not show a meaningful difference between the actual and required environmental condition, and five lacked essential information needed for reliable assessment. None was excluded because either system had performed well or poorly. Independent expert review established the underlying cause with strong evidence in 105 episodes (86.8%). In another 10 episodes (8.3%), the available evidence supported a likely cause but was not sufficient to confirm it conclusively. The remaining 6 episodes (5.0%) could not be resolved with confidence. Agreement among the three independent experts before consensus was high (Fleiss' $\kappa = 0.82$, 95% CI 0.77–0.87), indicating that the reference diagnosis was reproducible rather than dependent on a single expert's judgement. Among the 105 episodes with established causes, the learning AI-DT identified the correct cause as its first-ranked explanation in 81 episodes (77.1%), compared with 59 episodes (56.2%) for the frozen non-learning digital twin. The correct cause appeared among the first three explanations in 97 episodes (92.4%) for the learning AI-DT and 80 episodes (76.2%) for the frozen system. The learning advantage was therefore evident not only in whether the correct cause was considered, but also in how highly it was ranked.

The six unresolved episodes also revealed differences in how the two systems handled uncertainty. These episodes were retained in analyses of abnormality recognition, uncertainty and identification of additional information needed for diagnosis. They were not used to judge whether either system had identified the correct cause because the available evidence did not establish one with sufficient confidence. This prevented uncertain

reference evidence from being treated as a definitive answer. Corrective actions provided additional confirmation in 63 of the 105 episodes with established causes. In 52 episodes, correcting the suspected cause produced the expected improvement in indoor environmental conditions. In the remaining 11 episodes, the environmental response was incomplete or differed from what had been predicted, and further investigation was required before the cause could be established. These findings showed that an initially plausible explanation did not always represent the actual cause.

Overall, the reference-diagnosis results provided a consistent basis for comparing the two systems across the same real hospital problems. The learning AI-DT produced substantially more correct first-ranked diagnoses and more frequently placed the established cause among its leading explanations than the frozen non-learning digital twin. The findings therefore showed that learning from previously verified problem episodes improved diagnostic capability while preserving appropriate uncertainty where the available evidence remained insufficient.

Environmental Monitoring, Information Availability and Data Quality

Environmental monitoring provided sufficiently complete information for comparing the learning AI-DT with the frozen non-learning digital twin. Across the prospective study, 96.8% of expected continuous environmental measurements were available. Short gaps of five minutes or less accounted for 2.0% of observations, while longer missing periods accounted for 1.2%. No missing values were imputed in the main analysis. Allowing limited interpolation of short gaps in sensitivity analysis changed the estimated performance difference by less than 0.02 points. The learning AI-DT therefore remained clearly better than the frozen non-learning digital twin regardless of how these small data gaps were handled.

Data-quality checks identified 11 sensor-drift events, seven communication failures and four timestamp discrepancies. Three problem episodes contained enough measurement uncertainty to require separate sensitivity analysis. When these three episodes were excluded, the learning AI-DT still achieved an adjusted performance advantage of 0.62 points (95% CI 0.48–0.77) over the frozen non-learning digital twin, compared with 0.63 points in the full analysis. The better performance of the learning AI-DT was therefore not driven by a small number of episodes affected by poorer-quality measurements. Both systems also received the same additional information available for each problem. Building-management-system information contributed to 79 episodes, targeted airflow or pressure measurements to 57, maintenance or cleaning information to 39, structured information from occupants or hospital personnel to 31, and biological investigation to 14. The learning AI-DT continued to outperform the frozen non-learning digital twin across these information-rich episodes. Its advantage was particularly evident when several information sources had to be connected to explain a problem, showing that previous

learning improved how the available information was interpreted rather than simply increasing the amount of information available.

The 14 episodes requiring biological investigation provided a particularly demanding comparison. Eight were confirmed as involving moisture-related or microbial amplification, while three were ultimately shown to have non-biological causes despite reports of musty or unusual odours. Two had probable biological causes and one remained unresolved. Among the 11 episodes in which the cause was established, the learning AI-DT ranked the correct cause first in 9 episodes (81.8%), compared with 5 episodes (45.5%) for the frozen non-learning digital twin. This difference showed that the learning AI-DT was better able to distinguish a useful human observation from proof of the underlying cause. Audit checks found no instance in which information discovered later during field investigation was retrospectively added to either system's original assessment. The learning AI-DT therefore achieved its higher performance without access to information unavailable to the frozen system at the time of diagnosis. Together, these findings showed that its advantage remained after accounting for missing data, measurement uncertainty, different information sources and complex biological investigations, strengthening the conclusion that the observed improvement was associated with progressive learning from previously verified problem episodes.

Progressive Development of the AI-DT

The learning AI-DT and the frozen non-learning digital twin started the study with exactly the same computational knowledge of the hospital. This common starting point was called R_0 . From that point onward, however, the two systems developed differently. The learning AI-DT was updated after it gained verified experience from earlier problem episodes, while the frozen non-learning digital twin was never updated. By the end of the study, the learning AI-DT had progressed through 13 updates, from R_0 to R_{13} , whereas the frozen system remained at R_0 . By the time the learning AI-DT had reached R_{13} , its average performance score was 3.56 out of 4.00 for problems encountered during the final quarter, compared with 2.65 for the frozen R_0 system. This represented a 0.91-point performance advantage.

The records of these updates showed that the learning AI-DT was actually changing what it knew and how it approached later problems. As more verified problem episodes were encountered, it became better at judging which possible causes were more likely, ranking competing explanations, recognising relationships between hospital operations and environmental conditions, and deciding what additional information was needed when the cause was uncertain. The frozen non-learning digital twin could still analyse each new problem, but it could not use what had happened in previous problems to improve how it approached the next one. Among 30 final-quarter problem episodes in which the cause was established with strong evidence, the learning AI-DT in its later stage of development

culminating in R_{13} ranked the correct cause first in 26 cases (86.7%), compared with 18 cases (60.0%) for the frozen R_0 system.

The difference between the two systems became increasingly visible in their performance. During the first quarter, the learning AI-DT scored on average 0.28 points higher than the frozen non-learning digital twin. The difference increased to 0.54 points in the second quarter, 0.74 points in the third quarter, and 0.91 points in the fourth quarter. Statistical analysis confirmed that this increasing gap over time was unlikely to have occurred by chance ($P < 0.001$). This pattern is important because the learning AI-DT was not simply better from the beginning. The two systems began from the same starting point, but the learning AI-DT pulled progressively further ahead as it accumulated experience from previously verified problems. The frozen non-learning digital twin could not make the same progress because it continued to rely only on what it knew at the beginning of the study. The numerical difference between the later R_{13} stage and the unchanged R_0 system therefore reflected both better overall problem-solving performance and a substantially greater ability to place the correct cause first.

The archived computational records supported the same conclusion. They showed, step by step, how information from a verified problem contributed to learning and how that learning changed the AI-DT before it encountered later problems. In simple terms, the results showed that previous verified experience was being carried forward and used to improve future problem diagnosis and solving. Overall, the widening performance difference provided direct evidence of progressive learning. By the final quarter, the learning AI-DT's advantage over the frozen non-learning digital twin was more than three times larger than it had been in the first quarter. This showed that the learning AI-DT became increasingly capable as verified experience accumulated, whereas the frozen non-learning digital twin remained limited by the knowledge available at the start of the study.

Multidimensional Value-Oriented Performance

The independent evaluation of the two systems was highly consistent across the three expert assessors. Their overall agreement was 0.88 (95% CI 0.84–0.91), above the predefined acceptable level of 0.75. Agreement across the eight individual areas of performance was also strong, with weighted κ values ranging from 0.73 to 0.85. This gave confidence that the performance differences reported below reflected consistent expert judgement rather than large differences in how individual assessors interpreted the systems' outputs. The learning AI-DT performed better than the frozen non-learning digital twin across all eight areas of value-oriented problem diagnosis and solving. For problem framing and definition, the learning AI-DT achieved an average score of 3.32 out of 4.00, compared with 2.72 for the frozen system. For recognising that an abnormal indoor-environmental condition existed, the corresponding scores were 3.47 and 3.05. The smaller difference in abnormality recognition showed that the frozen system was already

reasonably capable of detecting when something was wrong, leaving less room for learning to improve this basic capability.

The largest advantages appeared in areas requiring deeper reasoning about why the problem was occurring. For distinguishing a symptom from its underlying cause, the learning AI-DT scored 3.06, compared with 2.32 for the frozen system. For connecting the problem with where it occurred, when it occurred and what was happening operationally at the time, the scores were 3.13 and 2.37, respectively. When generating and comparing several possible causes rather than accepting the first plausible explanation, the learning AI-DT scored 3.01, compared with 2.22 for the frozen system. When deciding what additional information was needed before reaching a reliable conclusion, it scored 3.29, compared with 2.51.

The learning advantage also extended from diagnosis to solving the problem. For generating and assessing possible solutions, the learning AI-DT scored 3.06, compared with 2.54 for the frozen non-learning digital twin. For value-oriented judgement, which considered whether a technically effective solution would also provide useful outcomes without unacceptable costs, disruption, safety consequences or other sacrifices, the learning AI-DT scored 3.02, compared with 2.59. Statistical tests that accounted for the eight simultaneous comparisons confirmed that the advantage of the learning AI-DT remained statistically supported in every performance area. The pattern of results therefore showed that learning contributed most where the system had to reason through competing explanations, connect information across place, time and hospital operation, and recognise what information was still missing. The improvement was smaller for simply recognising that an abnormal condition existed. This indicates that progressive learning strengthened the AI-DT's ability to move beyond identifying a problem towards understanding its likely cause and deciding what should be investigated or done next.

The learning AI-DT did not, however, outperform the frozen non-learning digital twin in every individual episode. It achieved a higher overall performance score in 94 of the 121 episodes (77.7%). The two systems achieved the same rounded overall score in 18 episodes (14.9%). These were mainly relatively straightforward problems for which the information available at the time was already sufficient for both systems to reach similar conclusions, leaving little additional benefit from previous learning. The frozen non-learning digital twin performed better in 9 episodes (7.4%). Four of these involved experience-related bias, where the learning AI-DT gave too much weight to a cause that had been important in previous problems but was not the best explanation for the current problem. The remaining five involved highly intermittent conditions, complex interactions between several ventilation zones, or several operational changes occurring at the same time. Under these conditions, previous experience did not provide a reliable advantage and sometimes directed the learning AI-DT towards an initially plausible but less appropriate explanation. These findings showed that learning improved performance overall, but

previous experience was not always beneficial and could occasionally mislead the learning AI-DT when a new problem differed importantly from problems encountered before.

Causal Diagnosis, Uncertainty and Additional Information Seeking

Among the 105 problem episodes in which the cause was established with strong evidence, the learning AI-DT identified the correct cause as its first-ranked explanation in 81 episodes (77.1%), compared with 59 episodes (56.2%) for the frozen non-learning digital twin. After accounting for similarities among problem episodes, the learning AI-DT was about 2.6 times more likely to place the correct cause first (OR 2.63, 95% CI 1.55–4.46, $P < 0.001$). The correct cause also appeared among the first three explanations in 97 episodes (92.4%) for the learning AI-DT, compared with 80 episodes (76.2%) for the frozen system. These findings showed that learning improved not only whether the correct cause was considered, but also how strongly it was prioritised over competing explanations.

Other measures of diagnostic performance showed the same pattern. The learning AI-DT correctly recognised the relevant causal category in 81% of cases where it was present, compared with 64% for the frozen system. Its ability to correctly rule out causes that were not responsible also improved from 82% to 88%. When the learning AI-DT identified a cause, that judgement was correct more often, with precision increasing from 68% for the frozen system to 80% for the learning system. Incorrect positive identification of causes fell from 18% to 12%. The learning AI-DT therefore became better both at finding genuine causes and at avoiding explanations that were not supported by the evidence.

This improvement involved only a small increase in computational time. The learning AI-DT took a median of 5.1 minutes to produce its locked assessment, compared with 4.7 minutes for the frozen non-learning digital twin. The additional 0.4 minutes was therefore accompanied by substantially better causal diagnosis rather than a large increase in the time required to reach an assessment. The learning AI-DT also became better at judging how confident it should be in its own conclusions. Its Brier score, which measures how closely expressed confidence matches what is later shown to be correct, improved from 0.214 for the frozen system to 0.149 for the learning system, with lower values indicating better performance. Expected calibration error similarly decreased from 0.13 to 0.06. In practical terms, the learning AI-DT was not only more often correct, but was also less likely to express high confidence when the available evidence did not justify it.

A further difference appeared when the information available was insufficient for a reliable diagnosis. The learning AI-DT requested additional information in 29 episodes, and 25 of these requests (86.2%) were subsequently shown to be useful for distinguishing among competing explanations. The frozen non-learning digital twin requested further information in 19 episodes, but only 11 requests (57.9%) were diagnostically useful. The learning AI-DT therefore became better at recognising what it did not yet know and what information

should be obtained next, rather than simply producing a diagnosis from whatever information was already available.

Value-Oriented Solution Assessment

The learning advantage extended beyond identifying what caused a problem. The learning AI-DT achieved an average value-oriented solution score of 3.02 out of 4.00, compared with 2.59 for the frozen non-learning digital twin. The adjusted improvement was 0.43 points (95% CI 0.27–0.59, Holm-adjusted $P < 0.001$). The learning AI-DT more consistently considered who was affected, what undesirable condition needed to change, what outcome should be achieved, the likely usefulness of a proposed solution, the resources and sacrifices required, hospital operating constraints and possible unintended consequences. The learning AI-DT was nevertheless not consistently strong in value-oriented solving. In 16 episodes, it proposed solutions that were technically plausible but did not adequately account for disruption to hospital operations, occupant comfort, workflow or another stakeholder sacrifice.

The same limitation occurred more often with the frozen non-learning digital twin, in 29 episodes. This difference showed that previous verified experience helped the learning AI-DT consider a broader range of consequences, but did not eliminate the problem completely. Safety-related trade-offs illustrated this limitation more clearly. In five episodes, the learning AI-DT proposed ventilation changes that could have conflicted with pressure-control or infection-control requirements if implemented without further review. The frozen non-learning digital twin produced comparable conflicts in 11 episodes. These proposals were not implemented automatically, but their occurrence showed why a technically effective solution could not be considered valuable solely because it might correct the immediate environmental problem.

Overall, the findings showed that better technical diagnosis did not automatically guarantee better problem solving. The learning AI-DT was better than the frozen system at both, but its improvement in value-oriented judgement was smaller than its improvement in several causal-reasoning dimensions. Learning therefore strengthened the ability to move from identifying a cause towards proposing a useful and acceptable solution, while leaving identifiable limits that still required professional review.

Contribution of Different Information Streams

The findings also showed which types of information helped the learning AI-DT outperform the frozen non-learning digital twin. This was examined by repeating completed analyses after removing one type of information at a time while leaving the remaining information unchanged. If performance became worse after a particular information source was removed, that source was considered to have contributed to the diagnosis. Removing the core environmental measurements produced the largest loss of performance. The learning AI-DT's overall score fell by 0.52 points, compared with a 0.43-point reduction for the

frozen non-learning digital twin. Removing building-management-system and operational information reduced performance by 0.41 points for the learning AI-DT and 0.30 points for the frozen system. Removing information describing spatial relationships and HVAC connections reduced performance by 0.36 and 0.24 points, respectively. These findings showed that both systems depended on measured environmental and building information, but the learning AI-DT made greater use of the relationships among these information sources.

Maintenance and cleaning information had a smaller overall effect. Removing it reduced the learning AI-DT's performance by 0.19 points, compared with 0.12 points for the frozen system. However, among episodes where maintenance or cleaning activities were later shown to be directly involved in the cause, removing this information reduced the learning AI-DT's score by 0.47 points. This showed that information that appeared unimportant overall could become critical for particular types of problems. Structured information from occupants and hospital personnel showed a similar pattern. Removing perceived-air-quality and other human-context information reduced overall performance by 0.15 points for the learning AI-DT and 0.10 points for the frozen system. For odour, thermal and intermittent location-specific problems, however, the learning AI-DT's performance fell by 0.39 points when this information was removed. Human observations were therefore especially valuable when the problem depended on where, when or under what activity a condition was experienced.

Taken together, these findings showed that the learning AI-DT did not become better simply by accumulating more sensor readings. Its advantage came increasingly from learning how to connect environmental measurements with building-system relationships, hospital operations, maintenance activities and human experience. The frozen non-learning digital twin received the same information for each problem, but could not use experience from previous verified episodes to improve how those different information streams were interpreted together.

Capability Boundaries, Failures, Safety and Robustness of the Findings

The learning AI-DT did not perform equally well under all types of problems. Among the 105 problem episodes in which the cause was established with strong evidence, 72 involved one dominant cause, while 33 involved several interacting causes. The learning AI-DT outperformed the frozen non-learning digital twin by an average of 0.78 points when one main cause could be identified, but by only 0.26 points when several causes interacted. This showed that previous learning was most useful when the new problem had a sufficiently clear relationship to earlier verified experience. Its advantage became smaller when several causes had to be disentangled at the same time. The same pattern appeared under different environmental conditions. In 68 episodes, the environmental signal was relatively stable or clearly distinguishable from normal background conditions. In these episodes, the learning AI-DT performed 0.71 points better than the frozen system.

In the remaining 37 episodes, environmental conditions varied considerably, making it more difficult to separate the problem from normal fluctuations. Under these conditions, the learning advantage fell to 0.29 points. The learning AI-DT therefore benefited more from previous experience when the environmental pattern of the new problem could be recognised clearly.

Where and how a problem spread through the hospital also affected the value of learning. In 64 episodes with relatively clear ventilation or physical pathways, the learning AI-DT outperformed the frozen non-learning digital twin by 0.76 points. In 41 episodes involving several ventilation zones or indirect pathways, the advantage decreased to 0.23 points. Previous experience therefore helped most when relationships between the source, affected locations and building systems could be connected clearly. Complex spatial interactions remained more difficult even after repeated learning. The timing of a problem produced an even stronger contrast. For 58 recurring episodes with patterns related to problems previously encountered, the learning AI-DT achieved an average advantage of 0.84 points. For 47 episodes involving intermittent, delayed or less familiar patterns, the advantage was only 0.21 points. This was one of the clearest indications that the benefit of learning depended on whether earlier verified experience was relevant to the new problem. Repeated experience was valuable, but it transferred less effectively when the timing or behaviour of the new problem differed substantially from what had been encountered before.

Hospital operating conditions produced a similar boundary. In 61 episodes where occupancy and HVAC operation were sufficiently clear to interpret, the learning AI-DT performed 0.73 points better than the frozen system. In 44 episodes where changes in occupancy, maintenance, hospital activities or HVAC operation occurred at the same time, the advantage fell to 0.18 points. The learning AI-DT therefore remained better overall, but simultaneous operational changes made it harder to determine which factor was actually responsible for the observed environmental problem. These repeated patterns showed where the learning AI-DT's capability remained strongest and where its advantage became limited. Progressive learning was most beneficial when new problems had clear causes, distinguishable environmental signals, understandable spatial pathways, recurring temporal patterns and interpretable operating conditions. Its advantage became much smaller when several of these relationships were uncertain or changing at the same time. These capability limits were based on repeated groups of episodes rather than isolated successes or failures.

The failure results reinforced the same conclusion. Among the 105 episodes with established causes, the learning AI-DT produced an incorrect high-confidence first-ranked diagnosis in 8 episodes, compared with 20 episodes for the frozen non-learning digital twin. The learning AI-DT therefore made substantially fewer confident diagnostic errors, but such errors were not eliminated. Its eight failures were concentrated in problems involving several ventilation zones, simultaneous operational changes or intermittent sources that

were no longer present when targeted measurements were taken. Four of the eight high-confidence failures also showed experience-related bias. In these cases, the learning AI-DT had encountered a similar cause previously and gave that familiar explanation too much importance when assessing the new problem. The explanation initially appeared plausible but was later shown to be incorrect. These failures demonstrated an important limitation of learning: previous experience could occasionally mislead the AI-DT when a new problem looked similar to an earlier one but had a different underlying cause. The frozen system could not develop this particular type of experience-related bias because it did not learn from earlier episodes, although its overall diagnostic performance remained poorer.

Safety-related findings showed a similar overall advantage but also confirmed the need for professional oversight. The learning AI-DT produced two recommendations that were judged potentially unsafe if implemented without further professional review, compared with five recommendations from the frozen non-learning digital twin. None of these recommendations was implemented. The learning system therefore produced fewer potentially unsafe recommendations, but learning did not remove the possibility of an inappropriate proposed action. The overall findings remained stable when the analysis was repeated under several more demanding conditions. When only the 105 episodes with strongly established causes were analysed, the learning AI-DT still outperformed the frozen system by 0.65 points (95% CI 0.50–0.80). When the three episodes affected by substantial sensor uncertainty were removed, the advantage remained 0.62 points (95% CI 0.48–0.77). Allowing limited estimation of very short data gaps changed the performance difference by less than 0.02 points. These results showed that the observed learning advantage did not depend on uncertain diagnoses, questionable measurements or the way small amounts of missing data were handled.

The advantage also remained when the influence of different hospital locations and problem categories was examined. Removing each problem category from the analysis one at a time produced adjusted performance differences ranging from 0.57 to 0.66 points. No single favourable category therefore explained the overall result. The learning AI-DT remained better than the frozen non-learning digital twin across different combinations of indoor-air, ventilation, thermal, acoustic, lighting and other indoor-environmental problems. Finally, the comparison remained focused on machine learning rather than deliberate improvement in professional human capability. Hospital professionals continued their normal work during RQ1, and four external professional-development activities relevant to facilities or environmental management occurred during the study. However, the AI-DT assessments remained hidden from the professionals managing the corresponding problems, and no research intervention was introduced to deliberately develop their problem-solving capability. The observed widening advantage of the learning AI-DT over the frozen non-learning digital twin was therefore not explained by a deliberate professional-learning intervention occurring alongside the machine-learning process.

Taken together, these findings showed that the learning AI-DT was more accurate, produced fewer high-confidence errors and fewer potentially unsafe recommendations, and retained its overall advantage under multiple sensitivity analyses. At the same time, its benefit was clearly conditional. Learning worked best when previous verified experience could be meaningfully connected to the new problem and became less useful, or occasionally misleading, when the new problem differed substantially from what had been learnt before.

Integrated Interpretation of the Findings

The findings supported the RQ1 alternative hypothesis ($H_{11}: P_{DT,I} > P_{DT,0}$) and rejected the null hypothesis ($H_{01}: P_{DT,I} \leq P_{DT,0}$), where $P_{DT,I}$ represents the performance of the progressively learning AI-DT and $P_{DT,0}$ represents the performance of the frozen non-learning digital twin that remained at the original R_0 representation. Across the prospective investigation, the progressively learning AI-DT achieved significantly greater value-oriented problem-diagnosis-and-solving performance than the frozen non-learning digital twin, and this advantage remained robust across the sensitivity analyses. The findings therefore showed that allowing the AI-DT to learn from previously investigated and independently verified real hospital problems produced meaningful improvement beyond continuing to use its original R_0 representation without learning. This finding is important for real hospital and building-industry practice because indoor-air and other indoor-environmental problems rarely present themselves with an obvious cause. Facilities teams may instead encounter complaints of odour, discomfort, poor air quality, pressure problems, temperature variation or other symptoms. The cause may occur elsewhere, depend on particular operating conditions or disappear before investigation begins. The learning AI-DT became increasingly better at combining environmental measurements, building-system information, hospital operations and human observations to determine what was likely to be wrong, what additional information was needed and what action should be considered.

The development of the AI-DT over time provided further evidence of progressive learning. Both systems began from the same R_0 , but only the learning AI-DT incorporated experience from previously verified problems. Its performance progressively separated from that of the frozen system as learning accumulated. The archived development from $R_0 \rightarrow R_1 \rightarrow \dots \rightarrow R_{13}$ showed that previous verified experience changed how the AI-DT ranked competing causes, connected different information sources and decided when further information was required. The diagnostic findings demonstrated the practical value of this development. The learning AI-DT was better than the frozen system at identifying and prioritising the actual cause, distinguishing symptoms from causes, comparing competing explanations, recognising uncertainty and requesting diagnostically useful additional information. This matters in practice because an incorrect first explanation can direct an investigation towards unnecessary measurements, inappropriate maintenance actions or solutions that address symptoms rather than underlying causes.

Learning also improved value-oriented solving. The learning AI-DT more consistently considered affected stakeholders, required outcomes, operational constraints, resources, sacrifices and possible unintended consequences when assessing solutions. This distinction is important because a technically effective intervention may still create problems for infection control, pressure relationships, thermal comfort, workflow, energy use or other hospital activities. Better technical diagnosis therefore did not automatically mean that every proposed solution would provide value. However, support for H_{11} did not mean universal superiority. Progressive learning was strongest for recurring problems with a dominant cause, distinguishable environmental signals, clear spatial relationships and interpretable operating conditions. Its advantage weakened when several causes interacted, environmental signals were highly variable, airflow pathways were complex, problems were intermittent or several hospital operations changed simultaneously. Experience-related bias also occurred when the AI-DT gave too much weight to a familiar explanation that had been correct previously but was inappropriate for the current problem.

The safety findings consequently reinforced the need for professional governance. Although the learning AI-DT produced fewer potentially unsafe recommendations than the frozen system, inappropriate recommendations were not eliminated. The AI-DT should therefore be understood as decision support rather than an autonomous authority, particularly where proposed actions affect ventilation, pressure control, infection control or other safety-critical building systems. The principal output of RQ1 was therefore a capability-boundary map of progressive AI-DT learning. It showed what improved as the AI-DT developed from R_0 to R_{13} , which information streams supported that development, where learning added substantial value, and where capability weakened, plateaued or became vulnerable to experience-related bias. For industry practice, the implication is that a learning AI-DT can help an organisation retain and reuse verified experience, so that recurring problems are not repeatedly approached as though nothing had been learnt from previous occurrences, while still requiring current evidence and professional judgement.

Overall, RQ1 showed that an AI-DT can progressively learn from independently verified real-world problem episodes and that this learning can improve value-oriented problem diagnosis and solving beyond the capability of its original non-learning representation. However, the improvement remained conditional on the quality and relevance of previous experience and did not eliminate uncertainty, failure or the need for professional governance. This provided the empirical foundation for RQ2, which examines whether deliberately developing professional human learning and enabling reciprocal professional human–AI-DT learning can produce capability beyond that achieved by the learning AI-DT alone.

Findings for Research Question 2:

Overview

Building on RQ1, which showed that the AI-DT could progressively learn from previously verified hospital problems, RQ2 examined whether value-oriented problem diagnosis and solving could be improved further when hospital professionals deliberately developed their own learning and worked reciprocally with the learning AI-DT. The results showed that coupling professional human capability with the learning AI-DT produced better performance than either the professional or the AI-DT working alone. Over the seven-month field programme, 24 hospital professionals participated and 23 completed the full study. A total of 103 possible problem episodes were identified, of which 88 met the predefined requirements for inclusion. Independent investigation established the cause with strong evidence in 77 episodes, identified a probable cause in seven, and left four unresolved.

The strongest overall performance occurred when the professional and learning AI-DT worked together. The coupled professional human–AI-DT system achieved an average score of 3.52 out of 4.00, compared with 3.17 for the AI-DT alone and 2.78 for the professional alone. After accounting for differences among professionals and problem episodes, the coupled system remained 0.35 points better than the AI-DT alone (95% CI 0.25–0.45, $P < 0.001$). The difference represented a substantial improvement (effect size 0.64). A separate statistical analysis similarly showed that the coupled system was more likely to achieve a higher performance rating (OR 2.28, 95% CI 1.74–2.98).

The results also showed that the benefit of coupling often went beyond simply combining the separate contributions of the professional and the AI-DT. In 58 of the 88 episodes (65.9%), the coupled system performed better than whichever of the professional or AI-DT had performed better independently. This stronger form of complementarity therefore occurred in almost two-thirds of the investigated problems. In another 20 episodes, coupled performance was approximately equal to that of the stronger independent contributor. In 10 episodes, however, working together reduced performance. The greatest gains occurred when the professional and AI-DT initially disagreed, when the problem contained substantial uncertainty, when several competing explanations had to be considered, or when stakeholder needs and operational consequences affected what would count as a valuable solution. Under these conditions, professional contextual knowledge and AI-DT analysis contributed different but relevant information and reasoning, allowing the combined assessment to become stronger than either independent assessment.

The benefit of coupling was smaller under severe time pressure, major information insufficiency, and situations in which the professional and AI-DT reinforced the same incorrect explanation. These findings showed that complementarity was substantial but not automatic or universal. Working together was most beneficial when the professional and AI-DT contributed different perspectives that challenged, corrected or extended one another's reasoning rather than merely agreeing. Overall, RQ2 extended the finding from RQ1. RQ1 showed that the AI-DT became more capable through progressive learning from

previously verified problems. RQ2 showed that value-oriented problem diagnosis and solving improved further when developing professional human capability was combined reciprocally with the learning AI-DT. The results therefore shifted the focus from machine learning alone to the additional capability that could emerge when professional human learning and AI-DT learning worked together.

Professional Participants, Intervention Exposure and Longitudinal Learning

The results showed that the professionals themselves became better at diagnosing and solving indoor-environmental problems as RQ2 progressed within the reciprocal professional human–AI-DT learning programme. Their improvement was evident in how they asked questions, reasoned through available information, formed judgements and considered what would provide value when deciding what should be done. Importantly, this development was also observed when their performance was assessed before they saw the AI-DT's assessment, showing that the improvement reflected developing professional capability rather than simply receiving better assistance from the AI-DT. All 24 professionals completed the baseline assessment, and 23 completed the midpoint and final assessments. They represented building services, HVAC engineering, facilities management, indoor-environmental or occupational-health functions and related technical roles with direct responsibility for indoor-environmental problem solving in the hospital.

Participation in the learning activities remained high throughout the programme. Among the 23 professionals who completed the study, attendance at facilitated learning activities was 94.1%, completion of episode-based reasoning records was 91.3%, and completion of structured reflections after the actual cause of a problem had been established was 89.6%. Three participants also attended external professional-development activities relevant to indoor-environmental problem solving, and five problem episodes involved substantial discussion across professional teams outside the structured intervention. These additional learning opportunities were retained in the study and later examined in sensitivity analyses rather than being ignored or removed.

The clearest evidence of professional learning came from new standardised problem cases that participants had not previously encountered. Their average score for the quality of questions asked increased from 2.18 out of 4.00 at baseline to 2.86 at midpoint and 3.31 at the end of the programme. Reasoning quality increased from 2.24 to 2.84 and 3.28, while judgement quality increased from 2.41 to 2.91 and 3.27. The quality of how participants represented their understanding of the problem increased from 2.20 to 2.79 and 3.24, and value-oriented reasoning increased from 2.32 to 2.97 and 3.38. All five areas showed statistically supported improvement over time (all $P < 0.001$), with large longitudinal effect sizes ranging from 0.91 to 1.24. The progressive improvement across previously unseen cases showed that professionals were not merely remembering solutions to earlier problems. They were developing ways of questioning, reasoning, judging and representing problems that could be transferred to new situations.

The improvement observed in the standardised assessment cases was also evident in the naturally occurring hospital problem episodes investigated during RQ2. Importantly, professional performance was assessed before the corresponding AI-DT output was shown to them. Their average independent performance increased from 2.55 during the first third of field episodes to 3.00 during the final third, with an adjusted increase of 0.39 points (95% CI 0.27–0.51, $P < 0.001$). This showed that professional capability itself improved during the programme. The improvement therefore could not be explained simply by professionals receiving better answers from the AI-DT at the moment of assessment.

The independent evaluators were also highly consistent when judging the professionals' observable reasoning. Their overall agreement was 0.86 (95% CI 0.81–0.90), while agreement across the individual assessment areas ranged from 0.72 to 0.84. This strengthened confidence that the observed improvement reflected a consistent change in what professionals demonstrated through their questions, reasoning, judgement and problem representation.

The results therefore provided evidence of professional human learning developing over time alongside continued AI-DT learning. The study did not claim to measure the professionals' internal mental models, M_t , directly. Instead, it assessed the observable evidence of how those mental models were developing, based on the questions professionals asked, how they reasoned, the judgements they formed and how they represented unfamiliar problems. This was important for RQ2 because the improved coupled professional human–AI-DT performance could then be understood as developing human capability working with developing machine capability, rather than a professional with unchanged capability simply receiving assistance from an increasingly capable AI-DT.

Prospective Problem Episodes and Independent Reference Diagnosis

Against the evidence above that the professionals themselves were developing their problem-solving capability, RQ2 also examined how this developing human capability performed when combined with the learning AI-DT across naturally occurring hospital problems. The 88 eligible problem episodes covered the same six broad types examined in RQ1: 20 particulate or gaseous indoor-air problems, 13 VOC or odour problems, 11 moisture or biological problems, 17 ventilation or pressure-control problems, 15 thermal-environment problems, and 12 acoustic or lighting problems. This ensured that the RQ2 findings reflected a broad range of indoor-environmental problems rather than one particular type. Of the 103 possible problem episodes initially identified, 15 were excluded using criteria established before the performance of the professional, AI-DT or coupled system was known. Five were continuations of an already registered unresolved problem, five did not ultimately satisfy the predefined requirements for a distinct problem episode, and five lacked essential information available at the time of the problem that was needed for a valid comparison. The exclusions were therefore based on the quality and

independence of the problem episode rather than on how well any of the three approaches performed.

Independent investigation established the cause with strong evidence in 77 of the 88 episodes (87.5%). A further seven episodes (8.0%) had a probable cause, while four (4.5%) remained unresolved. Agreement among the independent experts before they reached a final consensus was high (Fleiss' $\kappa = 0.84$, 95% CI 0.79–0.89). The probable and unresolved episodes were retained because they still provided useful evidence about problem framing, recognition of abnormal conditions, handling of uncertainty, information seeking and value-oriented judgement. Only analyses that required knowing the actual cause were restricted to the 77 episodes with strongly established causes. Audit checks also confirmed that information discovered later during the physical investigation was not added retrospectively to the original professional, AI-DT or coupled assessments. Each assessment therefore represented what could genuinely have been concluded from the information available at that time. This was important because the better performance of the coupled system could not be explained by it having access to information that became known only afterwards.

Value-Oriented Problem Diagnosis and Solving

The coupled professional human–AI-DT system performed better than the AI-DT alone across seven of the eight areas of value-oriented problem diagnosis and solving. Compared with the AI-DT working independently, coupling improved problem framing by 0.26 points, distinguishing symptoms from underlying causes by 0.33, connecting the problem with where, when and under what operating conditions it occurred by 0.31, and generating and comparing competing causes by 0.38 points. The coupled system also became better at deciding what additional information was still needed, improving by 0.46 points, and at generating solutions and considering their consequences, improving by 0.41 points. The largest improvement was in value-oriented judgement, at 0.52 points. All seven improvements remained statistically supported after adjustment for the multiple comparisons. Taken together, these seven areas represented a progression from understanding the problem, to determining why it was occurring, to deciding what further evidence was needed, and finally to judging what should be done. The improvement across this sequence showed that coupling strengthened the wider problem-diagnosis-and-solving process rather than improving only one isolated technical task.

The only area without a clear additional benefit from coupling was recognising that an abnormal environmental condition existed. The improvement was only 0.08 points and was not statistically supported after adjustment. This was consistent with the earlier RQ1 finding that the AI-DT was already relatively strong at detecting abnormal conditions. The added value of professional involvement therefore appeared mainly after a problem had been recognised, when the task became understanding why it was happening, what further information was required and what should be done about it. This distinction is important in

practice. Detecting that temperature, pressure, pollutant concentration or another environmental condition is abnormal is often only the beginning of an investigation. The more difficult task is deciding whether that abnormality is actually responsible for the reported problem, what else could explain it, and whether correcting it would solve the problem without creating another one. The results therefore located the main value of human–AI-DT complementarity in interpretation and judgement rather than basic detection.

The largest improvement in value-oriented judgement showed where professional contribution was particularly important. Professionals added knowledge about clinical workflow, operational feasibility, stakeholder needs and possible sacrifices that was sometimes absent or incompletely represented in the AI-DT's available information. Such knowledge can include practical circumstances that are difficult to represent fully in sensor or building-system data, such as whether a space can realistically be taken out of service, whether a proposed intervention would interrupt clinical work, whether access for maintenance is possible at the required time, or whether improving one environmental condition could create an unacceptable consequence elsewhere. These considerations determine whether a technically plausible intervention is also workable and valuable in the actual hospital.

Similarly, the strong improvement in identifying missing information reflected complementary strengths. The AI-DT could analyse patterns across large amounts of environmental and operational data, while professionals could recognise practical context and information gaps that were difficult to represent computationally. This suggests that the professional contribution was especially important at points where the problem could not be resolved simply by processing more of the information already available. The professional could recognise that a different question needed to be asked, that an additional observation or measurement was required, or that an apparently reasonable machine interpretation did not fit the operational reality of the hospital.

The overall pattern therefore revealed a useful division of strengths. The AI-DT contributed strongly to detecting and analysing patterns in large amounts of information, while the professional contribution became increasingly important as the task moved towards contextual interpretation, uncertainty reduction, consequence assessment and value judgement. The coupled advantage did not arise because the professional replaced the AI-DT's analysis, but because professional contextual knowledge extended what could be meaningfully inferred from that analysis.

The three independent evaluators were highly consistent in judging these outputs, with an overall agreement of 0.89 (95% CI 0.85–0.92). Their scoring also remained stable over the study. This strengthened confidence that the observed advantages of coupling reflected differences in the quality of the assessments rather than changes in how the evaluators applied the scoring criteria. The consistency of this pattern across independent evaluators was particularly important because several of the dimensions, such as problem framing,

consequence assessment and value-oriented judgement, required expert evaluation rather than being reducible to a single technical measurement. Stable agreement therefore strengthened confidence that the observed coupled advantage represented a reproducible difference in the quality of problem diagnosis and solving.

Causal Diagnosis, Uncertainty and Technical Performance

The advantage of coupling was also evident when the actual cause of the problem was known. Among the 77 episodes in which the cause was established with strong evidence, the professional working independently placed the correct cause first in 53 episodes (68.8%). The AI-DT working independently did so in 61 episodes (79.2%), while the coupled professional human–AI-DT system did so in 69 episodes (89.6%). After accounting for differences among professionals and problem episodes, the coupled system had more than twice the odds of placing the correct cause first compared with the AI-DT alone (OR 2.21, 95% CI 1.12–4.37, $P = 0.022$). This first-ranked result is important in practice because the explanation placed first is usually the one most likely to guide what is investigated, measured or acted upon next. A better first-ranked diagnosis can therefore reduce the likelihood of spending time and resources investigating a plausible but incorrect cause.

The same pattern remained when the first three possible explanations were considered. The correct cause appeared among the top three explanations in 87.0% of professional-alone assessments, 92.2% of AI-DT-alone assessments and 97.4% of coupled assessments. Incorrectly identifying a cause that was not actually responsible also became less common, with the false-positive rate decreasing from 0.15 for the professional alone and 0.11 for the AI-DT alone to 0.07 for the coupled system. Considering the top three explanations is also practically relevant because complex indoor-environmental investigations rarely begin with complete certainty. A useful diagnostic system should therefore not only identify one likely cause, but also keep the most credible alternatives visible for further investigation. The coupled system was better at doing both: it placed the actual cause nearer the top while reducing the number of unsupported explanations that could distract the investigation.

Coupling also improved how appropriately uncertainty was expressed. The Brier score decreased from 0.142 for the AI-DT alone to 0.105 for the coupled system, with a lower score indicating that expressed confidence corresponded more closely with what was subsequently found to be correct. In practical terms, the coupled system was not only more accurate but also better at knowing how much confidence should be placed in its conclusions. This distinction matters because a correct diagnosis expressed with unjustifiably low confidence may trigger unnecessary additional investigation, while an incorrect diagnosis expressed with excessive confidence may prematurely close the investigation. Better calibration therefore supported more appropriate decisions about whether to act, investigate further or keep alternative explanations open.

The combination of higher diagnostic accuracy, fewer false positives and better confidence calibration also suggests that coupling improved the quality of the diagnostic pathway rather than merely increasing the number of correct final answers. The coupled system was better able to prioritise the strongest explanation, suppress weaker unsupported explanations and communicate how certain that judgement should be. For facilities and hospital professionals, these are different but connected requirements for deciding what evidence should be sought next and how urgently action should be taken.

This improvement required additional time. The median time needed to produce a completed assessment was 10.8 minutes for the professional alone, 5.4 minutes for the AI-DT alone and 13.6 minutes for the coupled system. The coupled approach therefore provided better diagnostic and value-oriented performance, but it was not cost-free. Under routine conditions, the additional time was generally manageable, whereas under severe time pressure it became an important limitation. The practical trade-off was therefore between speed and the additional diagnostic value gained through interaction. The AI-DT alone produced the fastest assessment, while the coupled approach required time for the professional and AI-DT to compare interpretations, identify disagreement, consider missing information and reach a revised judgement. This additional time may be worthwhile for complex or consequential problems, but less appropriate where an immediate operational response is required.

This time finding also suggests that coupling need not be used in exactly the same way for every problem. Straightforward situations in which the AI-DT already has strong evidence may require little additional human–AI interaction, whereas ambiguous, high-consequence or value-sensitive problems may justify the extra time needed for deeper coupled reasoning. The results therefore point towards selective use of complementarity according to problem difficulty and decision consequence rather than assuming that maximum interaction is always desirable.

Taken together, these findings showed that the strongest performance did not come from the professional or the AI-DT simply replacing the other. It emerged when the developing professional capability and the learning AI-DT contributed different but useful information, reasoning and judgement to the same problem. The result was more accurate causal diagnosis, better handling of uncertainty and stronger value-oriented problem solving, while also revealing that the benefits of coupling depended on having sufficient time and meaningful complementary input from both sides. The central practical message is therefore that human–AI-DT complementarity improved not only whether the correct cause was found, but also how competing explanations were prioritised, how uncertainty was expressed and how efficiently further investigation could be directed. Its value lay in producing a more defensible diagnostic judgement, not simply a faster answer.

Evidence of Reciprocal Human–AI-DT Learning

The results showed that the interaction between the professionals and the AI-DT involved learning in both directions, rather than simply showing AI-DT outputs to professionals. In 62 of the 88 problem episodes, professional learning produced new or revised information that could be used by the coupled AI-DT. Across these episodes, 146 identifiable professional contributions were recorded. These included contextual relationships noticed by professionals, corrections or clarifications, possible causal explanations, questions that led to further information gathering, and considerations about stakeholders and the value of possible solutions. Importantly, professional contributions were not automatically treated as correct. Subsequent investigation supported 101 of the 146 contributions, contradicted 24, and left 21 unresolved. Only information that was later supported by independent evidence could contribute to subsequent machine learning. Of the supported contributions, 84 met the predefined requirements for incorporation into later AI-DT learning. This meant that professionals could provide information that influenced what the AI-DT considered, but they could not simply rewrite what the AI-DT had learnt. New professional information had to be independently verified before it could become part of later machine experience.

Learning also occurred in the opposite direction, from the AI-DT to the professional. In 71 episodes, the AI-DT identified at least one relevant relationship, explanation or information requirement that had not appeared in the professional's independent assessment. This new information changed subsequent professional reasoning in 51 episodes. In 34 episodes, it led the professional to ask another question, request an additional measurement or undertake further investigation. In 22 episodes, it changed how the professional ranked the possible causes of the problem. The AI-DT was therefore not merely providing an answer; its output frequently caused the professional to reconsider, investigate or revise the problem differently. The reciprocal effect was also visible from professional to AI-DT. Professional input materially changed the AI-DT's reasoning or the information it sought in 47 episodes. In 39 episodes, meaningful influence was observed in both directions: the AI-DT changed subsequent professional reasoning, and professional input also changed subsequent AI-DT reasoning or information seeking. These episodes provided the clearest evidence that the coupled system involved genuine reciprocal interaction rather than one-way assistance.

The longer-term effect of this reciprocity also became visible in the AI-DT itself. At the beginning of RQ2, the AI-DT that learnt only from its own verified problem experience and the AI-DT that could additionally learn from verified professional contributions started from exactly the same computational state reached at the end of RQ1. After seven update cycles, the AI-DT that had access to appropriately verified professional contributions performed better even before receiving professional input for the current problem. Its average score was 3.27 out of 4.00, compared with 3.17 for the AI-DT that continued learning without those professional contributions. The adjusted difference was 0.10 points (95% CI 0.04–0.16, $P = 0.002$). Although this additional machine-learning advantage was smaller than the immediate benefit obtained when the professional and AI-DT worked

together on the same problem, it was important. It showed that professional learning could contribute to later machine capability, not merely improve the assessment being discussed at that particular moment. In other words, information generated through professional reasoning in one problem could, after independent verification, become useful experience for the AI-DT when it encountered later problems.

The reciprocal process therefore developed over time rather than occurring all at once. A professional could learn from one problem and generate new information. After that information was independently verified, it could contribute to subsequent AI-DT learning. The updated AI-DT could then provide new information during a later problem, which could stimulate further professional questioning, reasoning and learning. The observed results were therefore consistent with the continuing pathway:

$$M_t \rightarrow \text{human learning} \rightarrow M_{t+1} \rightarrow \text{new/revised } I_m \rightarrow (I_m, R_t) \rightarrow \text{machine learning} \rightarrow R_{t+1} \rightarrow \text{new } I_h \rightarrow \text{further human learning} \rightarrow M_{t+2} \rightarrow \dots$$

In practical terms, the findings showed that the professional could help the AI-DT learn, and the AI-DT could help the professional learn, with verified experience from one problem contributing to stronger capability in later problems. This reciprocal development provided an important explanation for why the coupled professional human–AI-DT system outperformed either the professional or the AI-DT acting alone.

Conditions, Limits and Reliability of Human–AI-DT Complementarity

Further analysis helped explain why the coupled professional human–AI-DT system performed better, when the benefit was greatest, and where the approach remained limited. When previously verified professional contextual information was removed from completed coupled assessments, overall performance decreased by 0.18 points. Removing information obtained through professional-triggered clarification or additional measurements reduced performance by 0.16 points. Removing stakeholder and value information had a smaller effect on overall technical performance, reducing it by 0.09 points, but reduced value-oriented judgement by 0.28 points. Removing professional-generated causal hypotheses produced a smaller reduction of 0.08 points. These findings showed that professional contributions were most useful when they added relevant context, identified missing information or clarified what would provide value, rather than simply proposing additional possible causes. The key message was that professional contribution added value mainly by improving context, identifying what was still unknown and bringing stakeholder and practical considerations into the problem-solving process.

The benefit of coupling was also associated with observable improvement in how professionals approached problems. For every one-point improvement in the quality of the questions professionals asked, the coupled system's advantage over the AI-DT alone increased by 0.21 points (95% CI 0.12–0.30). A one-point improvement in reasoning quality was associated with a 0.18-point greater advantage (95% CI 0.09–0.27), while

clearer and more coherent external representation of the professional's understanding of the problem was associated with a 0.16-point advantage (95% CI 0.07–0.25). Further analysis indicated that improved questioning and reasoning accounted for approximately 40% of the observed coupled-system advantage, corresponding to an indirect pathway of 0.14 points (95% CI 0.08–0.21). These findings were consistent with the proposed reciprocal-learning mechanism, while not claiming that the professionals' unobservable internal thinking had been measured directly. The key message was that better professional questioning and reasoning were important observable mechanisms through which human–AI-DT coupling produced better performance.

The benefit of professional–AI-DT coupling was particularly strong when the two contributors initially brought different interpretations to the problem. In the 31 episodes where the professional and AI-DT initially disagreed about the leading explanation, coupling improved performance by 0.61 points over AI-DT alone. When their initial judgements broadly agreed, the improvement was only 0.20 points. Productive disagreement therefore created an opportunity for one contributor to question assumptions, introduce missing information or redirect the other's reasoning rather than simply confirm an existing view. The key message was that disagreement could be useful when it encouraged the professional and AI-DT to challenge and improve one another's reasoning.

A similar pattern appeared when the problem itself was difficult to interpret. Across 42 episodes with several plausible competing explanations, the coupled advantage was 0.48 points, compared with 0.24 points for problems with less diagnostic ambiguity. Where substantial stakeholder or operational trade-offs had to be considered, coupling improved performance by 0.56 points, compared with 0.20 points when such value considerations were relatively straightforward. Professional contextual knowledge was also particularly useful for multi-causal problems, simultaneous operational changes and complex spatial relationships. Coupling therefore helped reduce several capability limitations identified in RQ1 by bringing human contextual understanding into the AI-DT's analytical process. The key message was that coupling provided the greatest added value when problems were ambiguous, complex or required judgement about practical and stakeholder trade-offs.

However, complementarity weakened when the conditions for useful reasoning became too constrained. Under severe time pressure, represented by 18 episodes, the coupled advantage fell to 0.11 points and was not statistically distinguishable from zero. When available information was so incomplete that neither the professional nor the AI-DT could meaningfully distinguish among competing causes, coupling improved how the problem was framed and what additional information was requested, but did not reliably improve immediate causal diagnosis. Highly intermittent sources, rapidly changing environmental conditions and unresolved multi-zone airflow relationships also continued to create substantial uncertainty. The key message was that combining human and AI-DT capability

could not compensate for severe time pressure, insufficient evidence or environmental conditions that remained fundamentally difficult to interpret.

Coupling could also make performance worse. In 10 of the 88 episodes, the coupled system performed worse than whichever of the professional or AI-DT had performed better independently. Three failures involved automation bias, where professionals relied too heavily on an initially plausible AI-DT explanation. Two involved confirmation bias, where available information was interpreted in a way that reinforced an early assumption. Two involved unjustified rejection of correct AI-DT reasoning, another two involved reciprocal reinforcement of an incorrect explanation, and one involved excessive confidence despite insufficient evidence. These failures showed that reciprocal interaction could amplify error as well as correct it. Independent verification and tracking the origin of information prevented these incorrect conclusions from automatically becoming future learning examples. The key message was that reciprocity was beneficial only when human and machine reasoning critically examined one another; otherwise, they could reinforce the same error.

Safety findings showed the same need for continued professional governance. The coupled system produced two recommendations judged potentially unsafe if implemented without further review, compared with four from the AI-DT alone and three from the professional alone. None was implemented through the research system, and no adverse clinical or environmental event resulted from study-generated recommendations. The coupled system therefore reduced safety-relevant errors, but did not remove the need for independent evidence, professional judgement and authorised decision making. The key message was that coupling improved safety-related performance but did not justify autonomous implementation of AI-DT recommendations without professional review and governance.

The overall finding remained stable when the analyses were repeated under different assumptions. Restricting the comparison to the 77 episodes with strongly established causes produced a coupled advantage of 0.36 points (95% CI 0.26–0.46). Excluding episodes involving the participant who did not complete the programme produced an estimate of 0.34 points, while removing episodes potentially influenced by external professional training or substantial cross-team discussion produced 0.33 points. Alternative treatment of missing environmental information produced estimates between 0.33 and 0.36 points, while removing individual professionals, teams or problem categories one at a time produced estimates from 0.31 to 0.39 points. No single professional, team or problem type therefore explained the result. The key message was that the coupled-system advantage remained consistent even when potentially influential participants, episodes, missing data and problem categories were examined differently.

Overall, these findings showed that human–AI-DT complementarity was real but conditional. It was strongest when professional and machine capabilities contributed

different but relevant information, questioned one another's assumptions and helped resolve uncertainty or stakeholder trade-offs. It weakened when time, information or environmental conditions prevented meaningful reasoning, and it could occasionally amplify shared error. The final 88 independently adjudicated episodes exceeded the planned minimum of 80, providing the intended analytical basis for identifying not only whether coupling worked, but how, when and under what conditions it provided additional value beyond AI-DT learning alone. The key message was that professional human–AI-DT coupling created additional capability beyond AI-DT learning alone, but the value of that complementarity depended on the problem conditions and on the quality of the interaction between human and machine.

Integrated Interpretation of the Findings

The findings supported the RQ2 alternative hypothesis ($H_{12}: P_{h+DT} > P_{DT}$) and rejected the null hypothesis ($H_{02}: P_{h+DT} \leq P_{DT}$), where P_{h+DT} represents the performance of the coupled professional human–AI-DT system and P_{DT} represents the performance of the AI-DT acting alone. Across RQ2, the coupled professional human–AI-DT system achieved greater value-oriented problem-diagnosis-and-solving performance than the learning AI-DT acting alone. The adjusted advantage was 0.35 points (95% CI 0.25–0.45, $P < 0.001$), remained stable across sensitivity analyses and was not accompanied by an overall deterioration in safety-critical performance. The practical implication was therefore not simply that adding a professional improved AI-DT performance. Rather, developing professional capability and enabling reciprocal professional human–AI-DT learning produced additional capability beyond continued machine learning alone. This finding is important for hospital and building-industry practice because complex indoor-environmental problems are rarely resolved by data analysis or professional experience alone. The AI-DT could process large amounts of environmental, operational and building-system information, identify patterns and compare possible explanations. Professionals contributed contextual understanding of building operation, clinical workflow, stakeholder needs, practical constraints and information not adequately represented computationally. The strongest performance therefore emerged when these capabilities complemented rather than replaced one another.

Importantly, the professionals themselves also developed during RQ2. Their questioning, reasoning, judgement, problem representation and value-oriented thinking improved across previously unseen standardised cases and naturally occurring hospital problems. The coupled-system advantage therefore could not be interpreted simply as professionals receiving better answers from an increasingly capable AI-DT. Human capability was developing alongside machine capability. For industry practice, this is important because an effective human–AI system should support professionals in becoming better at questioning information, recognising uncertainty, forming defensible judgements and identifying when further investigation is required, rather than creating increasing dependence on machine outputs. The findings also showed that reciprocity operated in

both directions. Professional learning generated new or revised information that, after independent verification, could contribute to subsequent AI-DT learning. Conversely, AI-DT-generated information stimulated further professional questioning, reasoning, measurements and reconsideration of possible causes. Parallel AI-DT lineages further showed that verified professional contributions produced a measurable later machine advantage even before professional input was received for the current problem. Professional contribution therefore improved not only the immediate joint assessment but also, when verified, the machine capability available for later problems.

This reciprocal development gave practical meaning to $M_t \rightarrow M_{t+1}$ and $R_t \rightarrow R_{t+1}$. Professional learning changed how the human understood and approached subsequent problems, while machine learning changed the AI-DT's computational representation. New human information could become machine information, and new machine information could become human information. In practice, this creates the possibility of a continuing learning system in which verified organisational experience can move between professionals and the digital twin rather than remaining confined to either one. However, support for H_{12} did not mean that coupling was universally superior. Strong complementarity, $P_{h+DT} > \max(P_h, P_{DT})$, occurred in approximately two-thirds of episodes rather than all episodes. The greatest gains occurred when professional and AI-DT judgements initially disagreed, when several competing explanations existed, or when stakeholder and operational trade-offs were important. Under these conditions, different perspectives could challenge assumptions, expose missing information and improve judgement.

Complementarity weakened under severe time pressure, extreme information insufficiency and highly uncertain environmental conditions. In some episodes, interaction actually reduced performance. Automation bias occurred when professionals relied too heavily on an initially plausible AI-DT explanation, while confirmation bias and reciprocal reinforcement occurred when both sides strengthened the same incorrect interpretation. Human–AI interaction was therefore not inherently beneficial: effective interaction could correct error, whereas poor interaction could amplify it. The safety findings reinforced the need for professional governance. Coupling reduced potentially unsafe recommendations compared with either contributor acting alone, but did not eliminate them. Professional involvement should therefore not be assumed to make AI recommendations automatically safe, just as AI support does not make professional judgement infallible. Independent evidence, authorised decision making and professional oversight remain necessary for safety-critical hospital and building systems.

The principal scientific output of RQ2 was therefore a human–AI-DT complementarity and boundary map. It identified where deliberate professional learning added value beyond continued AI-DT learning, how human and machine development contributed to one another, which performance dimensions benefited most, and where reciprocal interaction weakened or failed. Overall, RQ2 showed that deliberately developing professional human

learning and coupling it reciprocally with an already learning AI-DT produced capability beyond AI-DT learning alone, but that this benefit remained conditional on relevant contributions, sufficient information and time, effective questioning and reasoning, and safeguards against shared error. The findings also exposed an important remaining limitation. The reciprocal system developed in RQ2 brought together professional knowledge and machine learning, but many indoor-environmental problems are experienced first by occupants, who may notice where, when and under what circumstances discomfort, odour, poor air quality or other problems occur. Such situated experience may not be fully captured by sensors, building-management systems or professional observations. RQ3 therefore became necessary to determine whether extending the reciprocal learning system from Professional Human–AI-DT to Occupant–Professional Human–AI-DT could provide additional capability beyond that already achieved in RQ2.

RQ3 consequently examined whether occupants could contribute meaningful information, questions, observations and value judgements to the developing system, whether participation could strengthen their own capability to understand and respond to indoor-environmental problems, and whether those contributions could improve value-oriented problem diagnosis and solving beyond the established P_{h+DT} capability. In this way, RQ3 extended the progression from machine learning in RQ1, to reciprocal professional human–machine learning in RQ2, and then to a broader reciprocal learning system that included the people directly experiencing the indoor environment.

Findings for Research Question 3:

Overview

Building on RQ2, RQ3 examined whether deliberately developing occupant learning, participation and agency could add further capability to the established professional human–AI-DT system. It therefore investigated participatory complementarity, treating occupants not simply as sources of complaints or environmental observations, but as participants whose situated experience, questions and judgements could contribute to value-oriented problem diagnosis and solving. The eight-month field programme was completed across eight hospital clusters. Eighty-four recurrent occupants entered the programme and 79 completed the final assessment. During prospective monitoring, 126 possible problem episodes were identified, of which 104 met the predefined requirements for inclusion, exceeding the planned minimum of 96. These comprised 45 comparison-period episodes and 59 intervention-period episodes. Independent investigation established the cause with strong evidence in 91 episodes, identified a probable cause in eight, and left five unresolved.

The main result showed that structured occupant participation improved the performance of the system already established in RQ2. The professional human–AI-DT system

achieved an adjusted average score of 3.47 out of 4.00, compared with 3.73 when occupants participated. The adjusted improvement was 0.26 points (95% CI 0.17–0.35, $P < 0.001$), and the occupant-participatory system was more likely to receive a higher performance rating (OR 2.14, 95% CI 1.62–2.83). The within-episode comparison supported the same conclusion. During intervention periods, performance improved by 0.31 points after structured occupant contributions were introduced, compared with only 0.07 points after a time-matched repeat review without structured occupant participation. The resulting difference-in-differences was 0.24 points (95% CI 0.15–0.33, $P < 0.001$), indicating that the improvement was not explained simply by reconsidering the same problem.

Occupants also became more capable participants over time, and their contributions measurably influenced professional reasoning, AI-DT processing and subsequent problem solving. The findings therefore supported the RQ3 alternative hypothesis (H_{13} : $P_{o+h+DT} > P_{h+DT}$) and rejected the null hypothesis (H_{03} : $P_{o+h+DT} \leq P_{h+DT}$), where P_{o+h+DT} represents the performance of the occupant–professional human–AI-DT system and P_{h+DT} represents the established professional human–AI-DT system. The benefit was nevertheless conditional rather than universal. Occupant participation added most value when occupants were familiar with the environment and when problems were perceptible, spatially specific or temporally patterned. It added less value when problems were difficult for occupants to perceive, information was severely incomplete, occupant accounts conflicted strongly, or subjective experience was given excessive weight over independently established technical evidence.

Occupant Participation, Retention and Intervention Exposure

The RQ3 programme included 84 recurrent occupants who regularly experienced the participating hospital environments: 50 healthcare workers, 19 patients with repeated exposure and 15 carers or visitors with recurrent presence. Seventy-nine completed the full longitudinal programme. Three participants left because they no longer had routine exposure to the participating area, while two patients could not complete final follow-up after discharge. No participant withdrew because of an adverse research experience. This mix of participants was important because indoor-environmental problems may be experienced differently by people who work continuously in a space, patients who occupy it repeatedly, and carers or visitors who encounter it under different conditions. Scientifically, this provided access to different forms of situated experience rather than relying on a single occupant group. In practice, it reflected the range of people whose comfort, health, activities and use of hospital spaces may be affected by indoor-environmental conditions.

Engagement with the intervention remained high. Among occupants exposed to the intervention, 95.2% completed the structured orientation, 91.8% used the structured participation process during at least one eligible problem episode, and 88.1% completed at least one reflection after the cause of a problem had been independently investigated.

Importantly, 61 occupants participated in two or more naturally occurring problem episodes. This provided repeated evidence of how occupant capability developed over time rather than relying only on one-off participation. The repeated involvement was scientifically important because learning is better demonstrated through change across multiple encounters than through a single response. For practice, it also showed that occupants could participate repeatedly in a structured way without the process being limited to one complaint or one isolated event.

The occupant records also showed a progressive improvement in how participants described and interpreted what they experienced. Early in the programme, observations, interpretations and possible causes were often combined into a single statement. Later contributions more clearly separated what the occupant directly noticed from what the occupant thought might explain it. For example, an occupant might report, *“the air feels stuffy only after the doors have remained closed”* as an observation, while separately proposing, *“the ventilation rate may be too low”* as a possible explanation. This distinction matters scientifically because an observation can be examined against measurements and other evidence, whereas a causal hypothesis must still be tested. Treating the two as equivalent could allow a plausible personal interpretation to be mistaken for an established cause.

This distinction was important because RQ3 was intended to develop occupants as more capable participants in problem diagnosis and solving, rather than simply collect more complaints. Within the occupant-learning framework, $I_{h,o}$ represented information received or experienced by the occupant, while $Exp_{h,o}$ represented what the occupant subsequently learnt through questioning, reasoning and reflection on that information. The repeated participation and increasingly structured contributions therefore provided early evidence that occupants were learning to contribute situated experience in a form that could be more meaningfully examined alongside professional and AI-DT reasoning. For industry practice, this changes the role of the occupant from a passive reporter of discomfort to a potentially useful participant in investigation. An occupant may know when a problem starts, where it is experienced, what activity was occurring, whether it is recurring, and what consequences it creates, even when that information is not visible in sensor or building-management data. However, RQ3 did not treat such experience as automatically correct. Its value depended on separating observation from interpretation and subjecting possible explanations to professional, AI-DT and independent physical evidence.

The broader scientific implication was therefore that occupant participation could become a structured source of learning rather than an uncontrolled source of subjective information. The practical implication was equally important: hospitals and other complex buildings may gain more value from occupants when they are supported to provide precise observations, ask useful questions and distinguish what they experienced from what they believe caused it. This creates a stronger basis for integrating occupant experience into value-oriented

problem diagnosis and solving without allowing subjective judgement to replace technical evidence.

Longitudinal Development of Occupant Capability

Occupant capability improved across standardised unfamiliar transfer cases administered at baseline, midpoint and post-intervention. The composite capability score increased from 2.03/4.00 at baseline to 2.70 at midpoint and 3.25 at post-intervention. The adjusted baseline-to-post increase was 1.19 points (95% CI 1.07–1.31, $P < 0.001$). The progressive increase from baseline to midpoint and then to the end of the programme was important because it showed development over time rather than a one-off improvement at the final assessment. Because the assessment cases were unfamiliar to the occupants, better performance also indicated that what they had learnt could be applied to new problems rather than simply remembered from previous episodes.

The largest improvement occurred in the ability to distinguish direct observation from causal interpretation. Mean performance on this dimension increased from 1.88 to 3.39. Relevance of questions increased from 1.96 to 3.18, interpretation of available information from 1.90 to 3.06, quality of judgement from 1.84 to 3.02, and articulation of useful outcomes and competing sacrifices from 2.11 to 3.46. Description of situated experience was comparatively stronger at baseline and increased from 2.28 to 3.42. Scientifically, this pattern showed that the intervention did more than help occupants describe what they experienced. They became better at separating evidence from explanation, asking questions that could advance an investigation, reasoning with available information and judging what outcome would actually provide value. The comparatively higher starting score for situated experience was also logical: occupants already knew what they experienced in the spaces they used, but initially had less capability to translate that experience into structured reasoning that could support diagnosis and solving.

These improvements were not confined to standardised cases. Among the 61 occupants participating in at least two real episodes, the mean capability score increased from 2.24 during their first eligible participation to 3.08 during later participation, an adjusted increase of 0.79 points (95% CI 0.65–0.93, $P < 0.001$). Improvement was therefore evident both under matched transfer conditions and in genuine hospital problems. This combination strengthened the interpretation of occupant learning. The standardised cases showed transfer under comparable assessment conditions, while the naturally occurring hospital episodes showed that the developing capability was also being used during actual indoor-environmental problems. In practice, this means occupant learning was not confined to an educational exercise; it became visible in how occupants participated when problems arose in the environments they genuinely experienced.

Three blinded evaluators assessed the externalised occupant responses. The overall ICC was 0.87 (95% CI 0.82–0.91), while weighted κ values across individual dimensions

ranged from 0.74 to 0.86. The results supported progressive change in observable questioning, interpretation, judgement and problem representation. They did not require the stronger claim that the internal mental model itself had been directly measured. This remained consistent with the methodological interpretation of $M_{t,o} \rightarrow M_{t+1,o}$ through repeated external evidence. In simple terms, the study could not look directly inside an occupant's mind and measure the mental model itself. What it could measure was how that developing understanding became visible through increasingly better questions, interpretations, judgements and representations of problems. The high agreement among independent evaluators strengthened confidence that these changes were observable and reproducible rather than dependent on one evaluator's opinion.

Prospective Problem Episodes and Independent Reference Diagnosis

The 104 eligible episodes represented six broad problem families. They included 23 particulate or gaseous indoor-air episodes, 16 VOC or odour episodes, 13 moisture or biological episodes, 18 ventilation or pressure-control episodes, 19 thermal-environment episodes and 15 acoustic or lighting episodes. The distribution therefore retained both indoor-air and other indoor-environmental problems. This breadth was important because the value of occupant participation could be examined across different kinds of problems rather than being inferred from one narrow category. It also reflected real building practice, where occupants may experience pollutants, odours, thermal discomfort, noise, lighting problems or pressure-related effects without initially knowing which technical system is responsible.

Twenty-two of the 126 registered candidate episodes were excluded using criteria fixed before the performance outcome was known. Nine were continuations of an already registered unresolved problem, seven did not ultimately satisfy the predefined requirement for a meaningful problem gap, and six lacked sufficient contemporaneous information for a valid prospective comparison. No episode was excluded because occupant participation appeared unhelpful. This was scientifically important because unsuccessful or unhelpful occupant contributions were allowed to remain visible in the dataset. The resulting findings therefore did not depend on selecting only episodes in which occupant participation appeared beneficial.

The reference panel classified 87.5% of eligible episodes as verified, 7.7% as probable and 4.8% as unresolved. Pre-consensus inter-rater agreement was high, with Fleiss' $\kappa = 0.83$ (95% CI 0.78–0.88). Among the 59 intervention-period episodes, 52 had verified causes, four were probable and three remained unresolved. Among the 45 comparison-period episodes, 39 were verified, four probable and two unresolved. The independent reference diagnosis was essential because occupant participation was being evaluated against evidence established separately from the occupants, professionals and AI-DT. A participant could therefore contribute useful experience without being assumed to know the true cause simply because that experience sounded plausible.

Occupant information did not function as its own reference standard. In 17 episodes, an occupant causal hypothesis was later contradicted by physical investigation. These hypotheses remained recorded as hypotheses and were not retrospectively converted into evidence. Conversely, 34 episodes contained occupant observations or contextual relationships that were independently supported and contributed materially to subsequent investigation. This preserved the distinction between situated experience, technical diagnosis and verified cause required by the methodology. This distinction has direct practical significance. Occupants can provide information that sensors or professionals may not capture, such as when an odour appears, which activity precedes discomfort or whether a condition occurs only in a particular location. However, an occupant's explanation of why it happened remains a hypothesis until independently tested. RQ3 therefore demonstrated a way to use occupant experience without confusing lived experience with technical proof.

Primary System-Level Performance

The principal cluster-randomised stepped-wedge analysis showed a reproducible improvement in P_{o+h+DT} compared with P_{h+DT} after adjustment for calendar period and clustered observations. The adjusted difference of 0.26 points remained statistically supported after accounting for repeated clusters, problem families, professionals and recurrent occupants. In practical terms, adding structured occupant participation improved the already-developed professional human–AI-DT system rather than merely adding another source of information. This was a demanding comparison because RQ2 had already established a relatively high-performing professional human–AI-DT system. The RQ3 result therefore represented additional capability beyond an already strong starting point.

The pattern of improvement differed across the eight established performance dimensions. Problem framing improved by an adjusted 0.18 points, symptom-cause discrimination by 0.15, spatial, temporal and operational reasoning by 0.27, generation and comparison of competing causes by 0.17, identification of missing information by 0.29, solution generation and consequence assessment by 0.31, and value-oriented judgement by 0.43 points. All seven effects remained statistically supported after Holm adjustment. The pattern showed where occupants added most value. Their contribution was particularly useful when the system needed to understand the circumstances surrounding a problem, recognise what information was missing, judge the consequences of possible actions and decide what outcome would actually be useful to those experiencing the environment.

Abnormality recognition improved by only 0.04 points and was not statistically supported after multiplicity adjustment. This was consistent with the already high capability of the professional human–AI-DT system established through RQ2. Occupants therefore contributed relatively little to recognising that an abnormal condition existed but contributed substantially more to explaining context, identifying missing information, evaluating

consequences and clarifying what constituted a useful solution. This has an important practice implication: occupant participation was most valuable after the basic technical question of “is something abnormal?” had largely been answered. Their stronger contribution was to questions such as “when does it happen?”, “who is affected?”, “under what circumstances?”, “what are we still missing?” and “would this proposed solution actually work for the people using the space?”

The evaluation panel maintained strong reliability, with an overall ICC of 0.90 (95% CI 0.87–0.93). Weighted κ values across the individual ordinal dimensions ranged from 0.76 to 0.87. The strong evaluator agreement was particularly important for dimensions such as value-oriented judgement and consequence assessment, which require expert judgement rather than a single instrument reading. It strengthened confidence that the observed occupant-related gains reflected consistent differences in assessment quality rather than inconsistent scoring.

Within-Episode Difference-in-Differences Analysis

The embedded within-episode comparison provided a separate test of whether improvement arose from occupant participation rather than simple repeated consideration. During the 59 intervention-period episodes, the first locked P_{h+DT} assessment had a mean score of 3.48, while the subsequent P_{o+h+DT} assessment increased to 3.79, producing an average ΔP_o of +0.31 points. This meant that, after the professional human–AI-DT system had already formed its assessment, introducing structured occupant information was followed by a further measurable improvement.

During the 45 comparison-period episodes, the equivalent first assessment averaged 3.45, while the time-matched second review without structured occupant participation averaged 3.52. The mean repeat-review change was therefore +0.07 points. This smaller improvement showed that simply giving the professional human–AI-DT system another opportunity to reconsider the problem could improve performance slightly. It therefore provided an important benchmark against which the additional contribution of occupant participation could be judged.

The adjusted difference between these changes was +0.24 points (95% CI 0.15–0.33, $P < 0.001$). The finding remained materially unchanged after adjustment for problem family, severity, location and diagnostic complexity. It therefore strengthened the interpretation that structured occupant participation added information or reasoning value beyond the benefit of simply looking at the problem again. This analysis was interpreted alongside, rather than independently from, the stepped-wedge result, as prespecified. The meaning is straightforward: both groups improved slightly when the problem was revisited, but the improvement was substantially larger when structured occupant participation was added. This made it less likely that the RQ3 effect was merely a “second-look” effect.

Taken together, these results established two connected findings. First, occupants themselves became more capable of contributing to problem diagnosis and solving through repeated learning, questioning and reflection. Second, when those developing capabilities were introduced into the established professional human–AI-DT system, system performance improved further. Scientifically, RQ3 therefore moved beyond asking whether occupant information was useful and showed that structured occupant learning could generate a developing source of situated knowledge that added measurable capability to an already learning human–AI-DT system. For practice, the implication is not that occupants should diagnose building problems independently, but that appropriately supported occupants can become informed participants whose observations, questions and value judgements help professionals and AI-DT systems understand problems that technical information alone may not fully reveal.

Technical Diagnosis, Uncertainty and Value-Oriented Solving

Among the 52 verified intervention-period episodes, the established professional human–AI-DT system ranked the verified cause first in 41 episodes (78.8%). After structured occupant participation, the verified cause was ranked first in 48 episodes (92.3%). The verified cause also appeared within the top three ranked explanations in 50 episodes (96.2%) after occupant participation, compared with 46 episodes before participation. This showed that occupant participation did more than add descriptions of discomfort or experience. In some cases, information about where, when or under what circumstances the problem was experienced helped the existing professional human–AI-DT system move the actual cause higher among competing explanations. For practice, this can help direct investigation towards the most relevant measurements, locations or operating conditions.

The false-positive rate decreased from 0.11 to 0.07, while the Brier score for uncertainty calibration improved from 0.128 to 0.095. These technical measures supported the broader multidimensional result but remained secondary to value-oriented problem diagnosis and solving. In practical terms, the occupant-participatory system became less likely to pursue an incorrect cause and better at matching its level of confidence to what was subsequently verified. This matters because excessive confidence in a wrong explanation can prematurely close an investigation, while appropriate uncertainty can signal that further evidence is still needed.

The greatest incremental effect occurred in value-oriented judgement. In 38 of 59 intervention episodes, occupants identified at least one stakeholder requirement, sacrifice or consequence that had not been adequately represented in the first locked professional human–AI-DT assessment. In 27 episodes, this information changed the ranking, design or acceptability of proposed solutions. Examples included sleep disturbance, privacy, inconvenience created by temporary relocation, disruption to clinical work and acceptability of local airflow changes. This finding showed why occupant participation could add value even when the technical cause was already reasonably understood. Professionals and the

AI-DT could identify what might technically solve the problem, while occupants could reveal whether that solution would create consequences that were important to the people actually using the space.

Occupant preference did not automatically override technical constraints. In seven episodes, a preferred occupant solution conflicted with infection control, pressure requirements or essential hospital operation. These preferences were retained as legitimate stakeholder evidence but were not adopted as technically acceptable solutions. The resulting assessments explicitly balanced stakeholder usefulness against non-negotiable safety and operational requirements. Scientifically, this preserved the distinction between understanding stakeholder value and allowing preference to determine technical truth. In practice, it showed that meaningful participation does not mean that every requested solution should be implemented; rather, occupant needs become part of a wider judgement that must also satisfy safety, engineering and operational requirements.

Demonstrated Occupant Agency

Agency was assessed from the prospective audit trail rather than from self-reported satisfaction. Across the 59 intervention-period episodes, structured occupants generated 286 distinct provenance-labelled contributions. Of these, 213 became new or revised I_h for the professional, 117 became new or revised I_m for the AI-DT, and 62 entered both pathways. Eighteen contributions did not enter either reasoning pathway because they were too nonspecific, duplicated information already available or were judged irrelevant. Agency was therefore demonstrated by whether an occupant contribution actually entered and influenced the problem-solving process, not simply by whether the occupant felt listened to. This provided a stronger scientific basis for claiming meaningful participation.

At least one occupant contribution demonstrably influenced the problem-solving pathway in 51 of 59 intervention episodes (86.4%). Occupant information generated a new measurement or targeted investigation in 30 episodes, altered problem framing or cause ranking in 23, influenced solution or consequence evaluation in 42, and introduced a previously overlooked stakeholder-value consideration in 37. These findings showed several ways in which occupant participation could change what happened next. An occupant contribution could redirect measurement, modify the interpretation of a problem, alter the proposed solution or reveal a consequence that had previously been overlooked. Agency therefore meant having an observable effect on reasoning or action, rather than merely supplying information.

Verified feedback was returned to participating occupants in 57 of the 59 intervention episodes. In the remaining two, patient discharge prevented direct post-episode feedback. Closing this feedback loop was important because occupants could compare what they initially observed or hypothesised with what subsequent investigation established.

Participation therefore became part of a learning cycle rather than a one-way process in which information was collected from occupants and never returned to them.

Reciprocal Occupant–Professional–AI-DT Learning

The audit trail showed that occupant contributions changed professional reasoning in 44 intervention episodes and changed AI-DT information processing, questioning or ranking in 31 episodes. Both pathways changed in 26 episodes. This demonstrated that occupant information was not merely stored alongside the technical assessment. It could enter professional reasoning, machine processing or both, depending on where the information was relevant.

The two RQ3 AI-DT lineages began from the same terminal RQ2 computational state. Across later prospective episodes, the lineage permitted to learn from appropriately verified occupant-derived information achieved a mean pre-current-occupant score of 3.59, compared with 3.52 for the lineage continuing without deliberate occupant-learning input. The adjusted difference was 0.07 points (95% CI 0.02–0.12, $P = 0.006$). Although smaller than the immediate benefit of occupant participation, this result was important because it showed that verified occupant-derived information could become useful machine experience for future problems. The value of occupant participation therefore did not necessarily end with the episode in which the information was provided.

The observed sequence was consistent with:

$(I_{h,o}, M_{t,o}) \rightarrow \text{occupant learning} \rightarrow M_{t+1,o} \rightarrow \text{occupant contribution} \rightarrow \text{new/revised } I_h, I_m, \text{ or both} \rightarrow \text{further professional or machine learning} \rightarrow \text{new } I_{h,o} \rightarrow \text{further occupant learning}.$

The timestamps showed that these developments occurred iteratively rather than simultaneously. In practical terms, an occupant could notice and learn something, contribute it to the system, influence professional or machine reasoning, and later receive new information that supported further occupant learning. RQ3 therefore extended reciprocity beyond professional and machine learning to include the people directly experiencing the indoor environment.

Mechanisms of Participatory Complementarity

Several recurring mechanisms explained why occupant participation improved performance. Previously unavailable spatial or temporal context occurred in 29 episodes, including localisation of draughts, recurrence of odours during specific activities and symptoms associated with particular operating periods. Diagnostically useful occupant questions contributed in 21 episodes, overlooked stakeholder requirements in 37, reconsideration of an initial professional or AI-DT explanation in 24, and disagreement between perceived conditions and sensor information triggered further investigation in 16 episodes. These mechanisms showed that occupants added most value where their

situated experience filled gaps that measurements, professional observation or existing AI-DT information could not fully represent.

A multilevel mediation analysis identified an indirect effect of 0.11 points (95% CI 0.06–0.17) through new diagnostic or stakeholder-value information generated by occupant participation, representing approximately 42% of the total adjusted intervention effect. This supported, without proving internal cognitive causation, the interpretation that much of the added value arose because occupant participation generated new information that changed subsequent reasoning. Scientifically, the result helps explain not only that participation worked, but how it worked. For practice, it suggests that occupant participation should be designed to elicit precise spatial, temporal, contextual and stakeholder information rather than simply invite general complaints or preferences.

Conditions, Limits and Reliability of Participatory Complementarity

The benefit of structured occupant participation depended strongly on how well occupants knew the environment and whether the problem could reasonably be experienced by them. Across 38 intervention episodes involving occupants highly familiar with the affected space, adding occupant participation improved performance by 0.34 points, compared with only 0.12 points across 21 episodes involving less familiar recurrent occupants. Problem perceptibility produced an even clearer difference. For 35 episodes involving conditions occupants could directly experience through odour, draught, noise, lighting, thermal sensation or recurring contextual patterns, the improvement was 0.38 points.

For 24 problems that were difficult to perceive without instruments, the gain was only 0.09 points. Occupant participation therefore added most value when people had relevant repeated experience that could reveal where, when and under what circumstances a problem occurred. The quality and consistency of occupant information were equally important. When occupants and professionals initially disagreed but subsequent investigation supported the occupant's contextual observation, performance improved by 0.46 points. In contrast, substantially conflicting occupant accounts produced an average gain of only 0.05 points. Differences between healthcare workers, patients and carers were comparatively small, with adjusted gains of 0.29, 0.22 and 0.24 points, respectively, and no statistically supported difference among groups ($P = 0.41$). The important factor was therefore not who the occupant was, but whether that person had relevant, sufficiently stable and useful situated experience.

The observed benefit was not explained simply by different types of problems occurring after the intervention or by occupants reporting more problems because they had become more attentive. Adjustment for problem type, severity, complexity and calendar period reduced the estimated effect only slightly, from 0.29 to 0.26 points. More importantly, when the analysis was restricted to 74 episodes identified independently through continuous monitoring, building-management systems or existing hospital reporting pathways,

occupant participation still produced an adjusted improvement of 0.23 points (95% CI 0.13–0.33). The benefit therefore remained even when occupant-generated reporting was removed as a possible explanation. Potential sharing of learning across hospital clusters also did not explain the result. Fourteen episodes were considered potentially exposed to cross-cluster spillover because professionals, discussions or recurrent occupants crossed between participating areas. Excluding these episodes produced an adjusted effect of 0.24 points, close to the primary result. This showed that the observed benefit was not dependent on unintended transfer of the intervention between clusters.

Participatory complementarity nevertheless had clear limits. Structured occupant participation improved performance by at least 0.20 points in 38 of the 59 intervention episodes, produced little material change in 13, and worsened performance in eight. Failures included misleading causal interpretations, excessive professional reliance on subjective experience, conflicting occupant accounts, dismissal of a valid observation, overreaction to a noticeable but technically unimportant experience, and reciprocal reinforcement of an incorrect explanation. Occupant participation could therefore correct blind spots, but it could also introduce or amplify error when subjective interpretation was treated as stronger evidence than it deserved. Safety findings reinforced this boundary. One occupant-participatory assessment contained a potentially inappropriate recommendation because a proposed comfort improvement conflicted with an established pressure-control requirement. Professional review identified the conflict and the recommendation was not implemented. No study-generated recommendation directly altered hospital systems, delayed clinical care or created or prolonged a harmful environmental condition. These failures were retained as genuine boundary conditions of participatory complementarity, rather than removed from the analysis.

Finally, the primary finding remained consistent across multiple sensitivity analyses. The adjusted effect was 0.27 points when restricted to the 91 verified episodes, 0.25 after excluding incomplete intervention exposure, 0.24 after excluding potential cross-cluster contamination, 0.23 among independently detected episodes, and 0.26 after adjustment for naturally occurring problem mix. The final dataset contained 104 independently adjudicated episodes, exceeding the planned threshold of approximately 96. Version-controlled records also preserved the connection between participant contributions, environmental information, professional reasoning, AI-DT states, verification and subsequent learning, with no unexplained breaks in provenance. Overall, participatory complementarity was reliable but conditional. It worked best when occupants had relevant familiarity and perceptible contextual experience, weakened when information was conflicting or when the problem was difficult for occupants to perceive without instruments or technical information, and required independent evidence and professional governance to prevent subjective experience from being mistaken for technical truth.

Integrated Interpretation of the Findings

The findings supported the RQ3 alternative hypothesis (H_{13} : $P_{o+h+DT} > P_{h+DT}$) and rejected the null hypothesis (H_{03} : $P_{o+h+DT} \leq P_{h+DT}$), where P_{o+h+DT} represents the performance of the occupant–professional human–AI-DT system and P_{h+DT} represents the established professional human–AI-DT system. Structured occupant participation produced an adjusted 0.26-point improvement (95% CI 0.17–0.35, $P < 0.001$). Seven of the eight performance dimensions improved, while abnormality recognition showed little additional benefit. The main contribution of occupants therefore occurred after a problem had been recognised, when the system needed to understand its context, reduce uncertainty, assess consequences and determine what would provide value.

The wider RQ3 proposition was also supported because the improvement was accompanied by evidence that occupants themselves developed capability and exercised agency. Over time, occupants became better at describing situated experience, separating observation from causal interpretation, asking relevant questions, interpreting information, forming defensible judgements and expressing desired outcomes and possible sacrifices. Their contributions also demonstrably changed professional reasoning, AI-DT information processing, investigation or solution assessment. Occupants were therefore not simply additional sensors or sources of complaints. They became progressively more capable participants in value-oriented problem diagnosis and solving.

This has important implications for hospital and building-industry practice. Occupants often experience aspects of environmental problems that may not be fully represented in sensor, BMS or professional information. They may know when an odour occurs, where a draught is strongest, whether discomfort follows a particular activity, how frequently a problem recurs, or how a proposed intervention would affect privacy, sleep, workflow or convenience. RQ3 showed that this situated experience can add value when it is structured, critically interpreted and examined alongside technical evidence rather than accepted automatically as fact. The findings also showed that occupant participation could contribute to learning elsewhere in the system. Occupant information could become new or revised professional information (I_h), machine information (I_m), or both. After appropriate verification, some occupant-derived information also contributed to subsequent AI-DT learning. The effect on later machine capability was smaller than the immediate participatory benefit, but it showed that occupant contribution could extend beyond the problem in which it was first provided. Verified experience from occupants could become part of the wider organisational learning process.

Support for H_{13} , however, remained conditional rather than universal. Occupant participation was most valuable when occupants were familiar with the environment, when the problem was perceptible through repeated experience, and when their contextual information helped resolve uncertainty or reveal stakeholder consequences. Participation added less when problems were difficult for occupants to perceive without instruments or technical information, when accounts conflicted substantially, or when subjective interpretation was given more weight than independently established evidence. This

distinction is important because meaningful participation does not mean treating occupant preference or explanation as technical truth. The findings also identified safety and governance boundaries. Occupant preferences occasionally conflicted with infection-control, pressure-control or essential operational requirements. Such preferences remained important stakeholder evidence, but they could not override non-negotiable technical and safety constraints. Occupant participation therefore worked best as part of a governed decision-support system in which experience, professional judgement, AI-DT analysis and independent evidence were considered together.

The principal scientific output of RQ3 was consequently an occupant empowerment, complementarity and boundary map. It established what occupants contributed beyond professionals and the AI-DT, whether occupants themselves developed capability and agency, how their contributions changed human or machine reasoning, how stakeholder usefulness and sacrifices affected solution evaluation, and under what conditions participation improved, failed to improve or weakened value-oriented problem diagnosis and solving. RQ3 therefore completed the cumulative progression from machine learning in RQ1, through reciprocal professional human–AI-DT learning in RQ2, to occupant-participatory reciprocal learning and agency in RQ3. The findings did not imply that occupants, professionals and the AI-DT learnt simultaneously or at the same rate. Rather, they showed that learning in one part of the system could generate new information that subsequently supported learning and problem solving elsewhere, creating an integrated human–AI-DT learning system in which verified experience could progressively contribute to better value-oriented decisions.

..... Chapter 5

When Mustafa completed his PhD, people around him saw the doctorate as the major achievement. Abdul Rahman saw the son who had taken the longest educational route in the family finally reaching the highest level of formal education. Mustafa had also become the first person in both his nuclear and extended family to earn a PhD. Mariam remembered the boy whose books had remained open late into the night while stronger examination results seemed to belong to other children. Safiya remembered how often Mustafa had quietly measured himself against her smoother academic journey, while Yusuf remembered the brother who had always been able to see a difficult problem differently even when his grades suggested nothing exceptional. Amina had seen something that the others had witnessed only in fragments: during the PhD, Mustafa had changed the way he responded to uncertainty, criticism, disagreement and his own mistakes.

The doctorate had not magically removed his weakness in externalising knowledge and understanding. Mustafa could still spend too long constructing an explanation, revise a paragraph repeatedly or struggle to reduce a complicated mental model to a few presentation slides. What had changed was his interpretation of those difficulties. He no longer allowed a weakness in expression to become immediate evidence that the thinking

behind it was weak, and he no longer regarded another person's fluency as automatic evidence of deeper intellectual capability. That distinction, which had taken him most of his life to discover, became increasingly important as he entered the next phase of his career.

The principle he carried from the PhD was simple to state but difficult to practise consistently: "The key life lesson this PhD taught me is that, no matter how much knowledge or experience we have, we should remain willing to question what we think we know. Verified experience strengthens judgement by distinguishing assumptions from evidence, refining understanding, and guiding better-informed decisions. We also grow when we allow AI and humans, including those whose voices may easily be overlooked, to improve our judgement, decisions and actions." Mustafa gradually realised that the significance of this statement would be determined not by how often he repeated it, but by what he did when his own judgement was the one being challenged.

He returned to academic life within a university whose identity centred on applied learning and applied research. The environment suited him better than many of the systems through which he had travelled earlier. Knowledge mattered, but so did what students and academics could do with it in authentic situations. Industry engagement mattered, practical consequences mattered, and problems did not always arrive already structured as examination questions. Mustafa found himself in an institution where his instinct to ask, "What problem are we actually trying to solve?" could be useful rather than inconvenient.

His early years as a faculty member were still difficult. Academic work demanded exactly the activities that had troubled him for decades: writing, presentations, proposals, committee discussions, promotion documents and public explanations of one's contribution. Mustafa could spend months developing an educational idea and then struggle to explain its significance in three paragraphs. A colleague might describe a smaller activity with remarkable clarity while Mustafa buried his central contribution beneath too much context. The old feeling occasionally returned, but now he recognised it before it became a verdict on himself.

Instead of concluding, "Perhaps my work is not good enough," he began asking a more disciplined question: "Is the work weak, or have I failed to externalise its value clearly enough?" Sometimes the work genuinely needed improvement. At other times, the contribution was strong but poorly communicated. The distinction allowed Mustafa to improve both without confusing them. He could revise an idea when the reasoning was weak and improve its communication when the reasoning was sound but inaccessible.

Artificial intelligence became useful precisely at this boundary. Mustafa used it to test alternative structures, expose gaps in an explanation, compress long arguments, generate questions that a sceptical reader might ask and examine whether the language he had produced actually represented what he meant. He never regarded the first AI output as the answer. Sometimes he rejected most of it; sometimes the machine noticed an ambiguity he

had overlooked; sometimes its criticism appeared persuasive until Mustafa traced the reasoning and found that it rested on an assumption that did not apply. AI reduced the friction of externalisation, but it did not remove his responsibility for what he accepted, rejected, revised or eventually communicated.

The human and machine learning processes he had studied during his PhD gave Mustafa a deeper way of understanding what was happening in these interactions. A human did not simply receive information and automatically become more knowledgeable. Information was encountered through an existing mental model; questions emerged from the interaction between that information and what the person already understood; thinking followed; learning could occur; and the resulting experience could refine the mental model brought to the next problem. A learning machine followed a different process: information entered an existing computational representation, generated queries for algorithmic processing, produced machine learning and experience, and, where the system was designed to learn progressively, refined the computational representation available for future problems.

More importantly for his academic career, Mustafa understood complementarity as an information-and-learning relationship rather than a simple division of labour in which the human “thinks” and the machine “writes”. AI-generated information could cause him to ask a question he had not previously considered and thereby contribute to human learning. His professional knowledge, questions and revised understanding could in turn improve the information provided to an AI-based system or reshape the problem it was being asked to examine. A colleague, student or practitioner could introduce contextual information that neither Mustafa nor the machine possessed. The important issue was therefore not who produced the first answer, but how the different contributions changed subsequent questioning, learning, judgement and action.

This understanding became valuable in scholarship. Before generative AI became available, Mustafa could sometimes spend several days knowing broadly what he wanted to argue but struggling to externalise the structure clearly enough to move forward. Now he might first establish the problem, his provisional position, the evidence supporting it and the assumptions that required scrutiny. AI could then provide alternative structures, challenge the logic, identify possible counterarguments or help him see that a sentence did not express what he thought it expressed. He would interrogate those outputs against the literature, his experience and the purpose of the work, while colleagues, students and practitioners supplied other perspectives that could expose what neither he nor the machine had seen.

That complementarity also influenced how he supervised research. When a student arrived with an AI-generated literature summary or a polished conceptual model, Mustafa rarely began by asking whether AI had been used. He asked what the student had learnt, which claims had been verified, which assumptions remained, why particular evidence had been

accepted, what alternative explanations had been considered and how the student's mental model had changed during the process. Conversely, he did not praise a student merely for avoiding AI. A student could write every word personally and still demonstrate weak reasoning, while another could use AI extensively yet remain responsible for a rigorous process of questioning, verification, judgement and revision.

This position increasingly gave coherence to Mustafa's research, teaching and educational leadership in an academic world still struggling to decide what human capability should mean in the age of AI. He did not want universities to replace one superficial measure with another. Before AI, polished externalisation could be mistaken for the whole of intellectual capability; after AI, the presence of machine assistance could be mistaken for the absence of human contribution. Mustafa considered both conclusions intellectually weak because they judged the final visible product without examining the process through which learning, judgement and decisions were formed.

That shift became important when Mustafa first encountered disappointment in academic promotion. By then, he had accumulated teaching innovations, publications, educational scholarship and contributions within the university, and he believed the case for progression was strong. The outcome was less favourable than he expected. Years earlier, such a decision might have reopened every old wound, from his secondary-school results to the university rejection and Second-Class Lower degree. For several days, the familiar feeling returned, but this time Mustafa did not allow it to remain unexamined.

He separated the event from the interpretation. The event was that he had not been promoted; his first interpretation was that the institution did not consider him sufficiently capable. He then asked whether the available evidence actually supported that broad conclusion. Feedback suggested something more specific: parts of his contribution were difficult for reviewers to see as a coherent programme of scholarship and impact. The problem was not merely that Mustafa had done too little. He had also failed to make sufficiently visible how his research, teaching, professional practice and educational leadership formed one connected ecosystem.

The distinction changed what he did next. He did not simply produce more outputs in the hope that quantity would eventually overwhelm doubt. He clarified the intellectual architecture of his work. His applied research in engineering education practice centred increasingly on developing communication and learning solutions that strengthened cognitive governance and lifelong learning capability, thereby providing upstream readiness for value-oriented problem diagnosis and solving. His teaching was not separate from that research, and his professional engagement was not an unrelated service activity. Each environment generated experience that informed the others.

Mustafa therefore began deliberately designing this ecosystem. In his Building Engineering and Management modules, students still learnt disciplinary knowledge, calculations,

systems and professional methods, but he increasingly structured learning around situations where knowledge had to be questioned and used. A student might receive information suggesting that a building was overheating and immediately recommend additional cooling. Mustafa would then introduce new information about radiant exposure, occupancy, air movement, operating schedules or user behaviour and ask whether the original judgement remained defensible. The purpose was not to make an ordinary problem artificially complicated; it was to let learners experience how quickly a plausible explanation could become an unexamined assumption.

He encouraged students to distinguish what they had observed from what they had inferred, what they knew from what they believed, and what further information would most efficiently reduce uncertainty. When students changed their minds after stronger evidence appeared, he deliberately treated the revision as intellectual progress rather than failure. A judgement was not valuable because it had been defended stubbornly. It was valuable because it could be justified by the best available evidence and revised when better evidence changed the situation.

He also designed learning so that students encountered their own reasoning before encountering AI. They might first record an individual interpretation of a case, then discuss it with peers, interrogate an AI system and finally revisit their original position. The interesting question was not whether the student's first answer matched the machine's answer. Mustafa asked what changed between the first and final judgement, why it changed, what information triggered the revision, which assumptions had been exposed and whether the final position was actually more defensible. In this way, AI became part of an iterative learning process rather than an answer-producing endpoint.

One classroom incident became particularly memorable. A group presented an AI-supported diagnosis confidently, and the slides were polished enough to impress the room. Mustafa asked a student near the back why the group had excluded another possible cause. The student hesitated and eventually admitted that the AI had ranked it as unlikely, so they had not investigated it further. Mustafa did not tell them they were wrong; he asked what evidence would be needed to justify excluding it and whether the AI's probability represented evidence from their actual case or a generalised inference.

The discussion changed immediately. Students who had been evaluating the quality of the presentation began examining the structure of the judgement. They discovered that the attractive answer depended on information that had never been verified for the case in front of them. Mustafa later used the episode in his research, not as proof that AI was dangerous, but as evidence of a learning-design problem worth investigating: how should education develop learners who can benefit from AI without surrendering the cognitive work required to govern its use? The incident became useful because it revealed a problem in thinking rather than merely a mistake in one answer.

Such questions became central to his scholarship. Mustafa was not conducting the same type of applied research as colleagues who developed new engineering technologies, materials, systems or industrial processes. His research was applied research in engineering education practice. He investigated how educational and communication solutions could be designed, refined and used to develop learners' capability to continue learning and solving problems when formal education could no longer provide a rehearsed answer. The scholarly contribution lay in understanding and designing for that development, not in claiming that education could pre-programme every future professional judgement.

This distinction became important within the university. Applied research for industry was part of the institution's signature identity, and many faculty members worked directly with companies to develop technologies and solutions. Mustafa's work supported that mission from another position. He asked what kind of human capability was required before, during and after such research if people were to identify the right problems, question assumptions, interpret evidence, learn from unsuccessful interventions and determine whether the resulting solution actually delivered value. His research therefore contributed to the human capability on which scientific research, applied research and professional practice depended.

This was where Mustafa saw his scholarship contributing to the university's wider applied-research ecosystem. His work aimed to develop learners and professionals who were better prepared to participate in scientific research, applied research and professional practice because they could ask better questions, recognise uncertainty, examine competing explanations, update mental models and make value-oriented decisions. He was not developing the new technology itself. He was working upstream on the capability of the people who would investigate, design, evaluate, adopt and improve such technologies.

Connecting that scholarly proposition to industry, however, proved much harder than Mustafa initially expected. University colleagues could understand why cognitive governance and lifelong learning might matter educationally, but a company facing an operational problem usually wanted something more immediate. It already had engineers, facility managers, consultants, contractors and technology vendors. Why should it trust a professor to guide experienced practitioners on how they ought to diagnose and solve problems? To some companies, an academic discussing problem framing and reflective judgement sounded dangerously close to someone who understood theory but had never carried responsibility for restoring a failed system before occupants complained, management demanded an answer and money began to be lost.

Mustafa discovered that an academic title could create credibility while simultaneously creating distance. Some industry contacts respected "Professor" but assumed that his contribution belonged in a classroom. Others were willing to invite him to speak for an hour

but were reluctant to let him near a live operational problem. A real consultancy engagement involved confidential information, commercial risk, professional liability, internal politics and decisions for which company staff remained accountable after Mustafa returned to the university. Trust could therefore not be demanded on the basis of publications, qualifications or university rank; it had to be accumulated through experience.

Several early approaches went nowhere. Emails received polite acknowledgements but no follow-up, and conversations ended with comments such as, “Interesting, perhaps we can explore this sometime.” Mustafa initially felt the old interpretation forming: perhaps industry did not see practical value in what he was doing. His post-PhD discipline made him examine that conclusion more carefully. The evidence showed only that companies were unwilling to commit significant trust before experiencing the value of his approach. That was a different problem, and therefore required a different solution.

..... Chapter 6

The university’s Continuing Education and Training school, or CET school, became an important bridge. The school already had mechanisms for engaging working professionals and companies, developing short courses, organising professional learning, handling enrolment and corporate arrangements, and providing a recognised university platform through which industry could interact with faculty without immediately entering a consultancy relationship. For Mustafa, this created a lower-risk environment in which professionals could experience his approach before deciding whether it had value for practice. He did not need to persuade a company at the outset to surrender a live problem to an academic; he could invite its professionals into a learning environment where authentic cases allowed them to examine how they themselves framed, diagnosed and solved problems.

His first few CET programmes were important not because they attracted large numbers of participants, but because they created repeated encounters between his scholarship and practitioners. Facility managers, engineers and other built-environment professionals came expecting technical updating. Mustafa gave them technical knowledge, but he also placed them inside cases where the obvious solution became less obvious as new information emerged. Some initially found the process uncomfortable. One experienced participant remarked that in real operations they did not have time to keep questioning everything because management wanted solutions.

Mustafa did not dismiss the comment. He asked the group to consider a different question: if an organisation implemented the wrong solution quickly, discovered two weeks later that the problem remained and then had to start again, had it actually saved time? The discussion that followed was more valuable than a lecture defending Mustafa’s position. Practitioners began supplying examples from their own experience: equipment replaced twice before the underlying control problem was discovered, repeated complaints

attributed to ventilation until an operational activity changed, and expensive interventions that solved one measurable problem while creating inconvenience elsewhere. Mustafa did not use these stories to prove that practitioners lacked critical thinking; he used them to show how operational pressure, incomplete information and established mental models could influence even highly experienced professionals.

The breakthrough into consultancy came gradually. A manager who had attended one of the CET programmes invited Mustafa to facilitate an internal session using a completed company problem rather than a live one. This mattered because the company could test his approach without exposing current operations to additional risk. Mustafa resisted the temptation to demonstrate expertise by telling the team what they should have done. He first asked them to reconstruct what information had been available at each stage, what they had believed at the time, why particular causes had seemed plausible and which later evidence had changed their understanding.

By the end of the session, the team could see several points at which an assumption had silently become treated as fact. They could also see that their final successful intervention did not prove that every earlier step in their reasoning had been sound. Some useful information had been ignored because it came from people considered peripheral to the technical investigation, while other information had been given excessive weight because it came from a familiar system indicator. Mustafa's value was not that he had arrived later with the correct answer; it was that he helped the organisation make its reasoning visible enough to learn from it.

Only after repeated interactions did the same organisation allow Mustafa to support a live problem. Even then, he was careful about his role. He was not replacing the company's engineers or assuming authority for decisions that properly belonged to authorised professionals. He was helping them govern the problem-solving process: making assumptions visible, examining the adequacy of evidence, identifying missing information, considering alternative explanations and evaluating whether proposed actions provided value to the affected stakeholders. Trust increased because his contribution could be experienced and evaluated rather than merely promised.

The CET platform consequently became part of Mustafa's research ecosystem rather than merely a channel for delivering courses. Undergraduate teaching allowed him to work with developing learners; postgraduate education exposed more advanced disciplinary reasoning; CET placed his ideas before professionals carrying years of situated experience; consultancy then tested whether the underlying principles remained useful when decisions had real operational, financial and human consequences. Differences among these environments became valuable research information. A learning activity that worked with students might appear simplistic to a senior engineer, while a professional shortcut that seemed efficient in routine work might become unsafe when transferred to an unfamiliar problem.

During one consultancy, an AI-enabled facilities platform repeatedly identified inadequate ventilation as the likely reason for complaints in a particular space. The recommendation was technically plausible, and several managers wanted to proceed with system modifications. During a site discussion, a junior technician quietly mentioned that the complaints appeared to follow a maintenance activity carried out every few weeks. The comment could easily have been ignored because the technician had no authority over the investigation, but Mustafa asked him to explain what he had observed.

The team reconstructed the timing of complaints, maintenance activities and system operation and discovered that the relationship deserved investigation. The AI had not been useless; it had analysed what it had been given. The technician had not proven the cause either; he had introduced information that the system did not contain. Mustafa's contribution was to create a process in which neither the machine nor the hierarchy decided prematurely what counted as relevant. The eventual diagnosis mattered professionally, but what Mustafa remembered was how easily the missing piece could have remained unheard.

That experience reinforced the sentence he had carried from the PhD: people grow when they allow AI and humans, including those whose voices may easily be overlooked, to improve their judgement, decisions and actions. It also gave practical meaning to complementarity. Machine analysis could reveal patterns across more data than any person could comfortably inspect; experienced professionals could contribute technical and operational context; technicians could provide situated observations; occupants could describe experiences that sensors did not capture; and Mustafa could help govern how these contributions were questioned, verified and integrated. No contribution became correct merely because of its source.

His consultancy practice gradually changed as a result. Mustafa became less interested in being the expert who arrived with the answer and more interested in leaving behind an organisation that could reason better after he had gone. He helped teams distinguish observations, interpretations, assumptions and verified findings. He encouraged them to record why a judgement had been made and what evidence would require it to be reconsidered. Successful interventions became sources of learning, but unsuccessful interventions became equally valuable when teams were willing to investigate why their reasoning had failed.

This professional work continuously strengthened his teaching. Real problems revealed cognitive difficulties that hypothetical classroom exercises sometimes missed. Professionals became anchored to familiar causes, ignored inconvenient evidence, confused fast action with good action or assumed that expensive technology must provide superior judgement. Mustafa translated such recurring cognitive barriers into cases and learning activities while protecting confidential details. Students therefore encountered patterns from real practice before facing them independently in industry.

His research ecosystem became increasingly iterative: professional practice generated problems; research investigated cognitive barriers and designed educational responses; curriculum allowed those responses to be implemented and refined; learner interaction produced new observations; professional education tested transfer beyond the classroom; and consultancy returned the ideas to real problems where their value and limitations became visible again. Human–machine complementarity became embedded within the same cycle. In teaching, AI could introduce alternative explanations that students had to interrogate; in scholarship, it could help Mustafa externalise and stress-test ideas; in CET, experienced professionals could expose where a technically plausible AI suggestion conflicted with situated operational knowledge; and in consultancy, machine analysis and human experience could be brought together without assuming that either source was automatically correct.

This became an important academic contribution in its own right. Mustafa increasingly argued that the educational question in the age of AI was not merely whether learners should be allowed to use machines. The more consequential question was whether the surrounding educational system was designed so that interaction with AI actually strengthened human learning and cognitive governance. If a student used AI to avoid questioning, interpretation and judgement, the technology could reduce learning. If the same student used AI to encounter alternative explanations, expose assumptions, identify missing information, test reasoning and refine a mental model, the same technology could support deeper learning.

The impact on learners gradually became visible in ways that mattered deeply to him. Former students began sending messages from industry. One described stopping a project team from replacing equipment because nobody had verified that the equipment was the source of the problem. Another explained how she had learnt to take occupant complaints seriously without automatically accepting the occupant's explanation of the cause. A postgraduate student told Mustafa that he had abandoned a research hypothesis after recognising that the available evidence could support association but not the causal claim he wanted to make.

His second attempt at academic progression therefore looked very different from the first. This time, Mustafa could explain how his teaching, research, professional engagement and educational leadership were connected. He could show not only outputs but the intellectual problem they addressed, the design principles guiding the work, how those principles had evolved through evidence and how the resulting solutions added value beyond his own classroom. When he was promoted to Associate Professor, he was pleased, but he did not experience the title as proof that the earlier rejection had been wrong. The earlier decision had exposed weaknesses in how his contribution was being developed and externalised, and he had learnt from it without allowing it to define his worth.

As his scholarship matured, Mustafa's work became increasingly recognised by educators, researchers and professionals beyond his institution. His conceptual work on cognitive governance and lifelong learning provided a language for capabilities that many people valued but struggled to describe. His communication solutions and fictional case stories invited learners into an iterative process of encountering information, questioning it, connecting it with existing mental models, reflecting on judgement and carrying the resulting way of thinking into another problem. The educational product was therefore not merely the final answer to a fictional case, but a learner who became more capable of recognising how he or she had reached a judgement and more willing to revise it when stronger evidence appeared.

Eventually, Mustafa was promoted to Professor of Engineering Education Practice for Building Engineering and Management. By then, the title represented a sustained body of scholarship, teaching innovation, industry engagement, professional education and leadership. It also represented something more difficult to quantify: learners and professionals who had encountered his work and subsequently approached problems differently. The promotion did not arrive as a dramatic reversal of everything that had happened before; it emerged from years of cumulative contribution becoming increasingly coherent, visible and useful.

His family understood the transformation in more personal ways. Amina had carried much of the household's financial burden while Mustafa pursued his PhD full time, and Mustafa refused to tell the later story as though success had resulted from his effort alone. As his academic and consultancy career strengthened, the couple became deliberate about ensuring that professional success translated into value for the family rather than simply more work. When their children struggled academically, Mustafa resisted the instinct to treat the grade as the central issue. He asked what they had understood, what had been misunderstood, what the feedback revealed and what they would try differently next time.

His children sometimes found these conversations irritating. On one occasion, one of them complained, "Dad, I only got one question wrong. Why are we analysing the whole thing?" Mustafa laughed and realised that cognitive governance could itself become burdensome when imposed without judgement. That small family moment taught him that even a valuable principle could become poor practice when applied without sensitivity to context. He apologised and let the matter go, recognising that reflective thinking itself required judgement about when reflection added value.

Abdul Rahman and Mariam, now older and retired, watched Mustafa's development with a mixture of pride and reflection. They remembered the boy whose educational journey had once caused them so much concern, the son who had needed three years of technical education and another three years of polytechnic before reaching university, and the young man whose persistence they had defended even when his results gave them little visible evidence of where that persistence might eventually lead. As Mustafa's career developed,

he and Amina made their later years more comfortable by supporting their healthcare, improving their living conditions, enabling them to travel and reducing financial concerns that had shaped parts of their earlier lives. Yet Mustafa increasingly understood that money addressed only one dimension of value.

Improving his parents' lives also meant giving them time, attention, companionship and presence. The relationship gradually changed from one in which Abdul Rahman and Mariam worried about Mustafa's future to one in which they increasingly relied on his judgement, support and presence. Mustafa listened when his father repeated stories from earlier years and invited Mariam to describe how she had experienced his difficult educational journey. When their memories differed, he became less interested in immediately deciding whose recollection was correct and more interested in what each perspective revealed and what could be learnt from the difference.

Safiya and Yusuf also became sources of intellectual companionship rather than comparison. Safiya's experience in business helped Mustafa understand organisational incentives, stakeholder behaviour and why technically rational recommendations could fail inside real institutions. Yusuf's computing expertise allowed them to debate AI, data and automation from different positions. Mustafa could admit that Yusuf understood something better than he did without feeling reduced by the admission, while Safiya could challenge his assumptions without reopening the old comparison between their educational journeys.

His respect in academia and industry grew partly because people noticed that Mustafa could carry expertise without making expertise oppressive. A junior engineer could challenge him, a student could expose an inconsistency in his explanation, an industry practitioner could reject one of his recommendations if the evidence justified doing so, and AI could produce an alternative that he had not considered. Mustafa increasingly saw all of these as opportunities to improve the quality of judgement. Authority did not require immunity from correction.

That did not mean he became endlessly uncertain or unwilling to decide. Cognitive governance was not paralysis. A decision often had to be made before complete information became available, particularly in professional practice where delay could itself have consequences. Mustafa taught learners and professionals to distinguish between waiting for impossible certainty and acting responsibly under uncertainty: what was sufficiently established to justify action, what remained uncertain, what consequences required monitoring and what future evidence should trigger reconsideration.

Over time, this became the connecting principle across his research, teaching, professional practice, educational leadership and everyday life. Scientific research required the willingness to revise an explanation when evidence contradicted it. Applied research required the ability to determine whether industry's stated problem was actually the problem worth solving. Professional practice required judgement under uncertainty,

education required learners who could continue developing after graduation, and everyday life required people capable of recognising when emotion, familiarity, hierarchy or convenience had become substitutes for evidence.

Years after his appointment, Mustafa sometimes remembered the primary-school classroom in which another pupil had quickly answered, “Plants need sunlight,” while he was still thinking about windows, changing light and mirrors. He no longer used that memory to prove that his younger self had secretly been the deeper thinker. That would reproduce the same ranking error in reverse. The more important lesson was that one visible performance could never reveal the entirety of either child’s developing capability.

Education had to assess what learners could demonstrate, but it also had a responsibility to develop what learners could eventually become. For Mustafa, that meant cultivating the capability to learn, question, think critically and reflectively, use logical deduction and abstract reasoning, imagine alternatives, form informed judgements and make decisions that guide value-oriented problem diagnosis and solving. Such capability provided upstream readiness not only for engineering practice but also for scientific research, applied research, education and everyday life.

The statement he had carried from his PhD therefore remained relevant long after the research itself had ended: “The key life lesson this PhD taught me is that, no matter how much knowledge or experience we have, we should remain willing to question what we think we know. Verified experience strengthens judgement by distinguishing assumptions from evidence, refining understanding, and guiding better-informed decisions. We also grow when we allow AI and humans, including those whose voices may easily be overlooked, to improve our judgement, decisions and actions.” By then, the statement had become less like a conclusion and more like a discipline through which Mustafa approached work and life.

By then, Mustafa understood that the most important word in the statement was not knowledge, experience, AI or even judgement. It was grow. Growth required him to continue questioning what he thought he knew, form judgements and make decisions from the best available evidence, act when necessary, and remain willing to revise his understanding, judgements and decisions when verified experience provided better evidence. The child who had once interpreted difficulty as proof of inadequacy had become a professor whose work centred on helping others remain capable of learning precisely when the answer was not yet clear. **The End!**