

ORIGINAL RESEARCH

Open access

A Theory-Building Model of Trustworthy Autonomous Smart-Grid Control Under Explainability, Operator Intervention, and Cyber-Physical Uncertainty

Reem Al-Harbi^{1*}, Omar Al-Qahtani², Salma Nasser¹

Abstract

Autonomous control is increasingly relevant to smart-grid operation, where control systems may support voltage regulation, frequency control, dispatch, and anomaly response. Yet their trustworthiness remains uncertain when decisions are opaque, intervention options are unclear, and cyber-physical conditions are unstable. Existing research often treats explainability, operator intervention, and cyber-physical uncertainty as separate design or evaluation issues. This separation limits theory development because trust in autonomous grid control depends on how these mechanisms operate together. This article develops a conceptual theory of trustworthy autonomous smart-grid control. The theory links explainability, operator intervention, and cyber-physical uncertainty to both perceived and operational trustworthiness. The article uses theory-building logic rather than empirical testing. It synthesizes concepts from trustworthy AI, human-automation interaction, explainable AI, and cyber-physical security to construct mechanisms and derive propositions. The proposed model specifies causal pathways through which explanation quality, intervention capability, and uncertainty handling shape trustworthiness. It also identifies interaction effects, boundary conditions, and testable propositions for future research. The theory provides a foundation for designing, evaluating, and empirically testing trustworthy autonomous control systems in smart-grid environments. It emphasizes that trustworthiness is not produced by transparency, human oversight, or robustness alone, but by their coordinated operation.

Keywords Theory-building, Explainability, Trustworthy autonomous control, Operator intervention, Cyber-physical uncertainty, Smart grid

*Correspondence:

Reem Al-Harbi
reem.harbi@gmail.com

¹ Department of Smart Electrical Systems, Faculty of Engineering, King Saud University, Riyadh, Saudi Arabia

² Department of Energy Data Systems, Faculty of Engineering, Qatar University, Doha, Qatar

Introduction

Autonomous smart-grid control is increasingly positioned as a response to growing operational complexity, renewable variability, cyber-physical interdependence, and the need for faster decision cycles in modern power systems. The idea of the autonomous grid reflects a movement from manual supervision toward increasingly intelligent control architectures, but this movement also raises questions about when operators should rely on algorithmic decisions and when they should challenge them [1]. In future control centers, human operators may no longer directly execute every control action, yet they will remain responsible for supervising, interpreting, and authorizing consequential system behavior [2]. This article therefore begins from the premise that autonomy in smart-grid control requires a theory of trustworthiness, not merely higher technical performance.

Trustworthy AI has been defined through multiple dimensions, including reliability, transparency, accountability, robustness, fairness, and governance [3]. In critical infrastructure, these dimensions become operational rather than abstract because autonomous control actions can affect system stability, customer service continuity, and resilience under disturbance. AI ethics frameworks emphasize that trust cannot be reduced to model accuracy, since socially and technically consequential systems must also be understandable, contestable, and accountable [4]. For autonomous grid control, this means that trustworthiness must be theorized as a relationship among system behavior, operator interpretation, and uncertainty management.

Explainability is central to this problem because operators need to understand why an autonomous system recommends or executes a control action before they can appropriately rely on it. General explainable AI research argues that explanations should make system reasoning intelligible to human users rather than merely expose technical details [5]. In smart-grid applications, explainability becomes more specific: operators may need to understand which measurements, constraints, contingencies, or cyber-physical anomalies shaped a control recommendation [6]. However, explanation alone may not be sufficient if operators lack effective authority to intervene when explanations reveal risk or inconsistency.

This article develops an original theory-building model of trustworthy autonomous smart-grid control under explainability, operator intervention, and cyber-physical uncertainty. The model proposes that trustworthiness emerges from three interacting mechanisms: the transparency of autonomous reasoning, the availability of meaningful intervention pathways, and the system's capacity to remain robust under noise, attack, and model mismatch. Human-automation research shows that appropriate reliance depends on calibrated trust rather than unconditional acceptance or rejection of automation [7]. Building on this insight, the article argues that trustworthy autonomous control is produced when explanations help operators understand actions, intervention mechanisms let them govern autonomy, and cyber-physical uncertainty handling preserves decision quality under degraded conditions.

Trust Problem in Autonomous Control

Defining trust in autonomous grid control

Trust in autonomous grid control can be defined as an operator's justified willingness to rely on an autonomous system for consequential control decisions under uncertainty. This definition emphasizes justification because trust should be grounded in evidence about system competence, transparency, and resilience rather than simple confidence in automation. Trustworthy AI scholarship distinguishes trustworthiness as a property of systems from trust as a human response, which is important because operators may overtrust weak systems or distrust strong systems [3]. In smart-grid control, the theoretical problem is therefore to explain how system properties produce calibrated reliance rather than mere acceptance.

Consequences of mistrust

Mistrust in autonomous control can lead operators to ignore useful recommendations, delay corrective action, or revert to manual routines that are too slow for cyber-physical operating conditions. Conversely, excessive trust can produce automation bias, in which operators accept algorithmic outputs despite incomplete explanations, warning signs, or anomalous system states [7]. The future role of operators in highly automated power systems is therefore not eliminated but transformed into a supervisory function requiring judgment about when autonomy should be trusted, constrained, or overridden [8]. The consequence of mistrust is not only reduced technology adoption but also degraded coordination between human expertise and machine speed.

The multi-dimensional nature of trustworthiness

Trustworthiness in autonomous smart-grid control is multi-dimensional because no single attribute can guarantee safe reliance across normal, uncertain, and adversarial conditions. Transparency supports interpretability, but transparent explanations may still describe actions based on corrupted data or uncertain models [9]. Robustness supports reliable performance, but robust control may still be rejected if operators cannot understand or contest it. Accountability and human oversight are therefore necessary complements to reliability and transparency, consistent with broader AI governance principles that treat trustworthy AI as a socio-technical property rather than a purely computational achievement [10].

Theoretical Foundations

Trust theories

Classical organizational trust theory explains trust through perceived ability, benevolence, and integrity, but autonomous smart-grid control requires these ideas to be translated into system-level terms. Ability corresponds to reliable control performance, integrity corresponds to rule-consistent and auditable behavior, and benevolence can be reinterpreted as alignment with operator goals, grid constraints, and public-service obligations. Contemporary trustworthy AI research extends this logic by emphasizing that technical systems must be designed to support justified human reliance through transparency, accountability, and robustness [3]. The proposed theory therefore treats trustworthiness as a latent construct that emerges when operators perceive that the autonomous system is competent, intelligible, controllable, and resilient.

Human-automation interaction theories

Human-automation interaction theory shows that trust is dynamic, context-sensitive, and shaped by experience with system performance, failures, and explanations. Endsley argues that lessons from automation research are essential for future autonomy because poorly designed autonomy can undermine situation awareness and encourage inappropriate reliance [7]. Human-autonomy teaming research further suggests that effective collaboration depends on communication, shared understanding, and calibrated expectations between human and autonomous agents [11]. In smart-grid control, these insights imply that autonomy should be theorized as a relationship of delegated authority rather than as a replacement for operators.

Theories of explainable AI and transparency

Explainable AI theory distinguishes between explanations that are technically accurate and explanations that are useful to human decision-makers. Miller argues that effective explanations are social and contrastive because users typically want to know why one action was taken rather than another [9]. Arrieta and colleagues similarly frame explainability as a bridge between model complexity and human understanding, especially when AI systems affect high-stakes decisions [5]. For smart-grid control, transparency must therefore be operationally meaningful, showing not only model logic but also the grid conditions, uncertainty sources, and constraints that shaped a control recommendation.

Cyber-physical security and resilience theories

Cyber-physical security theory emphasizes that autonomous grid control operates in environments where measurement streams, communication channels, and control commands may be noisy, delayed, manipulated, or incomplete. False data injection research demonstrates that corrupted measurements can mislead state estimation and downstream decision-making, making cyber-physical uncertainty directly relevant to trustworthiness [12]. Reviews of smart-grid cyber threats show that attacks can target sensing, communication, control, and market layers, producing uncertainty that is both technical and organizational [13]. The theory developed here therefore treats uncertainty handling as a trust mechanism because operators cannot justifiably rely on autonomy that is transparent and controllable but fragile under adversarial or degraded conditions.

Explainability Mechanism

Definition and forms of explainability

In this theory, explainability refers to the autonomous control system's capacity to present the reasons, evidence, constraints, and uncertainty conditions behind its recommendations or actions in a form operators can use. Explainability may include feature-level explanations, rule-based summaries, counterfactual comparisons, confidence information, causal narratives, or operational rationales. XAI research in energy and power systems highlights that explanations must be adapted to domain-specific decisions, such as voltage regulation, load forecasting, congestion management, intrusion detection, and dispatch [6]. The mechanism operates by converting opaque algorithmic behavior into interpretable operational meaning.

Effects on operator understanding

The model proposes that explainability increases operator understanding when it clarifies how system inputs, constraints, and objectives are connected to control outputs. In smart-grid contexts, interpretable machine learning can help operators identify which grid variables, topology conditions, or temporal patterns influenced a prediction or recommendation [14]. Understanding is not equivalent to agreement, because an operator may understand a recommendation and still reject it if it violates local knowledge or risk tolerance. However, without understanding, operators cannot distinguish between appropriate reliance, cautious monitoring, and necessary intervention.

Effects on perceived trustworthiness

Explainability is hypothesized to increase perceived trustworthiness by making autonomous control appear more competent, accountable, and aligned with operational goals. Explanations can reduce the psychological distance between machine decision logic and operator judgment, especially when they show why a specific action is preferable under current grid conditions [15]. In smart-grid stability and control settings, explainable approaches can make predictive or control outputs more acceptable by revealing the factors that support system recommendations [16]. The theory therefore treats explainability as a primary pathway from technical autonomy to perceived trustworthiness.

Boundary conditions

The effect of explainability on trustworthiness is bounded by explanation quality, operator expertise, workload, and the reliability of the underlying information. Poor explanations can produce confusion, false reassurance, or misplaced trust, while overly detailed explanations can overload operators during time-sensitive events. Explainable intrusion detection research in smart grids illustrates that explanations must support rapid interpretation of security-relevant alerts rather than simply expose model internals [17]. Thus, explainability improves trust only when it is timely, decision-relevant, faithful to system behavior, and matched to the operator's control task.

Operator Intervention Mechanism

Modes of intervention (override, veto)

Operator intervention refers to the formal capability of human operators to pause, modify, veto, override, or approve autonomous control actions. In highly automated power systems, intervention may occur before execution through approval gates, during execution through override commands, or after execution through rollback and corrective control [8]. The autonomous grid literature implies that stronger autonomy does not remove the need for human authority, especially when actions affect safety, reliability, or public-service obligations [1]. In this theory, intervention mechanisms convert human oversight from a symbolic governance claim into an operational control function.

Antecedents of effective intervention

Effective intervention depends on timely alerts, clear explanations, usable interfaces, operator competence, and well-defined authority boundaries. Human-autonomy teaming research shows that coordination improves when human and autonomous agents share task understanding and communicate relevant information at the right level of detail [11]. Training also matters because operators must learn when to rely on autonomy, when to question it, and how to intervene without destabilizing the system [18]. The model therefore proposes that intervention capability is not merely the existence of an override button but the presence of conditions that make intervention informed, timely, and legitimate.

Effects on trust and autonomy acceptance

Operator intervention mechanisms are hypothesized to increase trust by preserving human agency under uncertainty and reducing fear of uncontrollable automation. When operators know that they can inspect, veto, or override autonomous actions, they may be more willing to accept autonomy during routine operating conditions. Studies of human-autonomy teaming suggest that trust calibration improves when users understand both the system's capabilities and their own role in supervising it [19]. In smart-grid control, this means that well-designed intervention mechanisms may increase autonomy acceptance without encouraging blind reliance.

Interaction with explainability

The operator intervention mechanism interacts with explainability because interventions are more effective when operators understand why the autonomous system recommends a given action. An explanation that identifies the triggering condition, expected system benefit, and uncertainty level can help the operator decide whether to approve, modify, or block the action. Research on future power system control centers emphasizes that operators will need decision-support environments that integrate automation outputs with human judgment rather than separate them [2]. The theory therefore proposes that explainability and intervention jointly produce governable autonomy: explanations make action intelligible, while intervention makes autonomy controllable.

Cyber-Physical Uncertainty Mechanism

Types of cyber-physical uncertainty

Cyber-physical uncertainty refers to conditions that make autonomous smart-grid control decisions less reliable because system inputs, models, communication channels, or operating assumptions are incomplete, unstable, or corrupted. In this theory, uncertainty includes stochastic sensor noise, missing or delayed measurements, forecast error, topology mismatch, parameter drift, communication latency, and adversarial manipulation. Surveys of false data injection in smart grids show that compromised measurements can distort the information base on which control systems depend [20]. More broadly, uncertainty handling in cyber-physical systems requires attention to both technical sources of error and the operating context in which autonomous decisions are made [21].

Impact on autonomous decision quality

Cyber-physical uncertainty affects autonomous decision quality by weakening the correspondence between the system's internal representation and the actual grid state. When measurement streams are noisy or manipulated, an autonomous controller may select actions that appear optimal within its model but are unsafe, inefficient, or inappropriate in the physical system. False data injection and related smart-grid attacks can alter state estimation, mislead anomaly detection, and degrade downstream control decisions [22]. The model therefore proposes that uncertainty reduces trustworthiness not only by increasing error probability, but also by making system behavior harder for operators to interpret and validate.

Moderating role of transparency and intervention

The theory proposes that transparency and intervention moderate the negative effect of cyber-physical uncertainty on trustworthiness. Transparency helps operators see whether a recommendation is based on stable evidence, uncertain

forecasts, suspected anomalies, or degraded data quality, while intervention allows them to slow, constrain, or reverse autonomous action when uncertainty becomes operationally unacceptable. Reviews of cybersecurity in cyber-physical power systems emphasize that attack detection and defense must be connected to operational response rather than treated as isolated security functions [23]. Thus, uncertainty handling becomes trustworthy when the system exposes uncertainty clearly and gives operators meaningful authority to act on that exposure.

Trustworthiness under adversarial conditions

Under adversarial conditions, trustworthiness requires more than normal-case performance because the autonomous system must remain interpretable, governable, and robust when its information environment is actively manipulated. Cyber threats to smart grids can affect sensing, communication, control, and coordination layers, making autonomous decisions vulnerable to hidden distortion [13]. Reviews of cyber-physical attacks on future energy systems further show that attacks may interact with distributed energy resources, digital control, and increasingly automated infrastructure [24]. The model therefore hypothesizes that trustworthiness under attack depends on whether the system can identify suspicious conditions, communicate their implications, and shift from autonomous optimization to supervised mitigation.

Theory Model

Integrating the three mechanisms

The proposed theory integrates explainability, operator intervention, and cyber-physical uncertainty handling as three interdependent mechanisms that jointly shape trustworthy autonomous smart-grid control. Explainability makes autonomous reasoning visible, intervention makes autonomous action governable, and uncertainty handling makes autonomous control resilient under degraded or adversarial conditions. Trustworthy AI frameworks justify this integration because they treat transparency, accountability, reliability, and robustness as mutually reinforcing dimensions rather than independent features [3]. The model therefore proposes that trustworthiness emerges most strongly when the system is simultaneously understandable, controllable, and resilient.

Causal structure and relationships

The causal structure of the model begins with autonomous control recommendations and moves through three trust-generating pathways. First, explanation quality increases operator understanding, which strengthens perceived trustworthiness when explanations are faithful, timely, and relevant to operational decisions [9]. Second, intervention capability increases perceived control and accountability, which supports calibrated acceptance of autonomy in human-automation settings [7]. Third, cyber-physical uncertainty handling preserves decision quality by detecting degraded conditions, adapting control confidence, and supporting secure response under false data or attack scenarios [12].

Model boundary and assumptions

The model applies to autonomous or semi-autonomous smart-grid control systems that support voltage regulation, frequency control, congestion management, dispatch, anomaly response, or related operational decisions. It assumes that human operators remain institutionally accountable for system oversight even when autonomy executes or recommends control actions. It also assumes that explanations are intended for trained operators rather than general consumers, which aligns with energy-system XAI research emphasizing domain-specific interpretability [6]. The model does not claim that trustworthiness can be achieved through interface design alone, because cyber-physical resilience and governance authority remain necessary boundary conditions.

Model diagram (described)

The conceptual diagram of the theory would show autonomous smart-grid control at the center of a left-to-right causal architecture, with explainability, operator intervention, and cyber-physical uncertainty handling represented as three

parallel mechanisms feeding into perceived and operational trustworthiness. Explainability would connect to operator understanding, intervention would connect to perceived control and accountability, and uncertainty handling would connect to robustness and decision quality. Interaction arrows would show that explainability strengthens intervention quality, intervention constrains autonomous action under uncertainty, and uncertainty disclosure changes how explanations are interpreted [2]. This diagram would visually express the model's central claim: trustworthy autonomous control is a coordinated socio-technical condition rather than a single technical attribute.

Figure 1 presents the proposed theory model in which explainability, operator intervention, and cyber-physical uncertainty handling jointly transform autonomous smart-grid control from opaque technical automation into perceived and operational trustworthiness.

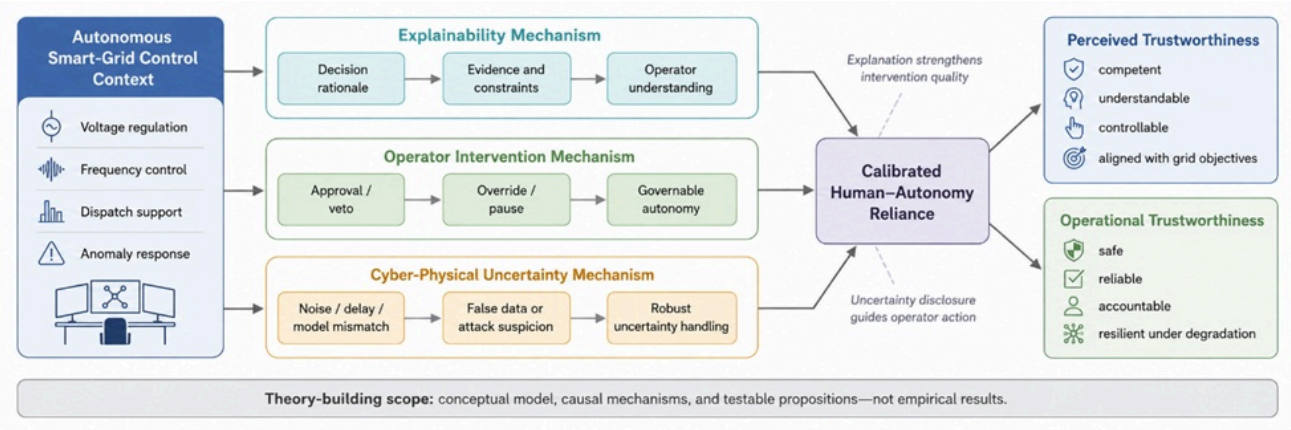


Figure 1. Integrated Theory Model of Trustworthy Autonomous Smart-Grid Control under Explainability, Operator Intervention, and Cyber-Physical Uncertainty

Table 1 outlines the key constructs, underlying mechanisms, and proposed relationships that shape operator trust and effective control in autonomous smart-grid systems.

Table 1. Theoretical Model: Constructs and Hypothesized Relationships

Core construct	Conceptual definition	Primary mechanism	Hypothesized relationship
Autonomous smart-grid control	A control system that recommends or executes grid actions with partial independence from direct human command	Delegated decision authority	Greater autonomy increases the need for explicit trust-generating mechanisms because operators must rely on system reasoning without manually controlling every action
Explainability	The system's ability to communicate reasons, evidence, constraints, and uncertainty behind control decisions	Cognitive transparency	Higher explanation quality increases operator understanding, which increases perceived trustworthiness when explanations are faithful and operationally relevant
Operator understanding	The operator's ability to interpret why an autonomous action is recommended and what grid condition it addresses	Sensemaking	Understanding mediates the relationship between explainability and calibrated trust because operators cannot rely appropriately on actions they cannot interpret

Operator intervention capability	The ability to approve, pause, veto, modify, override, or reverse autonomous control actions	Governable autonomy	Stronger intervention capability increases autonomy acceptance by preserving human agency and accountability
Intervention effectiveness	The degree to which intervention is timely, informed, usable, and institutionally legitimate	Human-on-the-loop control	Intervention improves trustworthiness only when operators receive enough information and authority to act before autonomous decisions become operationally irreversible
Cyber-physical uncertainty	Noise, delay, model mismatch, forecast error, topology error, or adversarial manipulation affecting control decisions	Robustness under degradation	Higher uncertainty weakens autonomous decision quality and perceived trustworthiness unless detected, disclosed, and managed
Uncertainty transparency	The system's ability to communicate uncertainty type, severity, and operational implication	Risk disclosure	Uncertainty transparency moderates trust by helping operators distinguish reliable recommendations from decisions requiring caution or intervention
Cyber-physical robustness	The system's ability to maintain acceptable control behavior under degraded or adversarial conditions	Resilient decision quality	Robustness strengthens operational trustworthiness by reducing the probability that autonomy converts corrupted inputs into harmful actions
Perceived trustworthiness	The operator's justified belief that the system is competent, understandable, controllable, and aligned with grid objectives	Calibrated reliance	Perceived trustworthiness increases when explainability, intervention, and uncertainty handling mutually reinforce one another
Operational trustworthiness	The system's actual capacity to support safe, reliable, accountable, and resilient control decisions	Socio-technical assurance	Operational trustworthiness is highest when technical robustness is coupled with human oversight and interpretable decision logic

Propositions

Propositions on explainability–trust link

The model proposes three propositions about the explainability–trust relationship. Proposition 1: higher explanation quality will increase operator understanding of autonomous control decisions. Proposition 2: operator understanding will mediate the positive relationship between explanation quality and perceived trustworthiness, consistent with XAI theories that treat explanation as a bridge between complex models and human judgment [5]. Proposition 3: explanation usefulness will have a stronger effect on trustworthiness than explanation detail because operators need actionable operational meaning rather than excessive technical disclosure.

Propositions on intervention–trust link

The model proposes that operator intervention capability has both direct and indirect effects on trustworthiness. Proposition 4: stronger override, veto, and approval mechanisms will increase perceived trustworthiness by preserving operator agency and accountability. Proposition 5: intervention effectiveness will mediate the relationship between formal intervention capability and autonomy acceptance because a nominal override option is insufficient unless it can be used

under real operating constraints [18]. Proposition 6: excessive or poorly timed intervention demands will reduce trust by increasing workload, uncertainty, and responsibility without improving control quality.

Propositions on uncertainty moderation

The model proposes that cyber-physical uncertainty moderates the effects of both explainability and intervention on trustworthiness. Proposition 7: cyber-physical uncertainty will weaken the relationship between autonomous control performance and perceived trustworthiness when uncertainty is hidden or poorly communicated. Proposition 8: uncertainty transparency will reduce this weakening effect by helping operators interpret whether a recommendation is based on reliable evidence, degraded data, or suspected manipulation [21]. Proposition 9: cyber-physical robustness will strengthen operational trustworthiness by limiting the effect of attacks, noise, and model mismatch on control decisions [23].

Propositions on interaction effects

The model's final proposition concerns the interaction among all three mechanisms. Proposition 10: trustworthy autonomous smart-grid control will be highest when explainability, operator intervention, and cyber-physical uncertainty handling are jointly present and mutually reinforcing. This proposition follows from human-autonomy teaming research, which suggests that trust calibration depends on both system communication and the human role in collaborative control [9]. It also reflects smart-grid cybersecurity research showing that operational resilience requires coordinated detection, interpretation, and response rather than isolated technical safeguards [13].

Implications

Implications for system design

The theory implies that autonomous smart-grid control systems should be designed as explainable, governable, and uncertainty-aware decision environments rather than as isolated optimization engines. Designers should connect each control recommendation to a concise operational explanation, a visible uncertainty status, and a clearly defined intervention pathway. Explainable AI for smart-grid intrusion detection shows that explanations are most valuable when they help operators interpret abnormal conditions and decide what action to take [17]. Therefore, system design should integrate explanation panels, confidence and uncertainty indicators, override logic, escalation triggers, and audit records into a unified supervisory architecture.

Implications for future research

Future research can test the model by examining how explanation quality, intervention capability, and uncertainty transparency affect operator reliance, response time, perceived control, and decision quality in simulated or field-based smart-grid control settings. Studies can also compare whether different explanation forms, such as counterfactual explanations, feature attributions, alert rationales, or uncertainty narratives, produce different levels of calibrated trust. Emerging work on interpretable decentralized smart grids suggests that explainability can be evaluated in relation to both local control decisions and system-level stability [25]. Further theoretical development should also examine institutional accountability, regulatory assurance, and cross-organizational trust in increasingly autonomous energy infrastructures.

Conclusion

This article developed a theory-building model of trustworthy autonomous smart-grid control under explainability, operator intervention, and cyber-physical uncertainty. The central claim is that trustworthiness does not emerge from autonomous performance alone. It emerges when operators can understand autonomous reasoning, intervene meaningfully in autonomous action, and rely on systems that remain robust under uncertain and adversarial conditions.

The model contributes a conceptual structure for explaining how trust is generated, weakened, and calibrated in autonomous grid-control environments. It distinguishes perceived trustworthiness from operational trustworthiness while showing how both are shaped by the same interacting mechanisms. This distinction is important because a system may appear trustworthy without being resilient, or it may be technically resilient without being accepted by operators.

The theory also clarifies why explainability, intervention, and uncertainty handling should not be designed as separate modules. Explanations make intervention informed, intervention makes autonomy governable, and uncertainty handling determines whether explanations and interventions remain meaningful under degraded conditions. Trustworthy autonomous control is therefore a coordinated socio-technical achievement.

Future empirical work can operationalize the constructs and test the propositions in simulation, control-room studies, digital twins, and field deployments. Such research can examine when trust is appropriately calibrated, when automation bias emerges, and when operators reject or overrule autonomous decisions. The long-term value of the theory lies in guiding autonomous smart-grid systems that are not only intelligent, but also understandable, governable, resilient, and worthy of justified reliance.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 04 Apr 2025

Revised: 16 May 2025

Accepted: 28 Jun 2025

Published online: 20 September 2025

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

Biagini V, Subasic M, Oudalov A, Kreusel J. The autonomous grid: Automation, intelligence and the future of power systems. *Energy Res Soc Sci*. 2020;65:101460.

Marot A, Kelly A, Naglic M, Barbesant V, Cremer J, Stefanov A, et al. Perspectives on future power system control centers for energy transition. *J Modern Power Syst Clean Energy*. 2022;10(2):328-44.

Thiebes S, Lins S, Sunyaev A. Trustworthy artificial intelligence. *Electron Mark*. 2021;31(2):447-64.

Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell*. 2019;1(9):389-99.

Arrieta AB, Díaz-Rodríguez N, Del Ser J, Benetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. 2020;58:82-115.

Machlev R, Heistrene L, Perl M, Levy KY, Belikov J, Mannor S, et al. Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. *Energy AI*. 2022;9:100169.

Endsley MR. From here to autonomy: lessons learned from human–automation research. *Hum Factors*. 2017;59(1):5-27.

Prostejovsky AM, Brosinsky C, Heussen K, Westermann D, Kreusel J, Marinelli M. The future role of human operators in highly automated electric power systems. *Electr Power Syst Res*. 2019;175:105883.

Miller T. Explanation in artificial intelligence: Insights from the social sciences. *Artif Intell*. 2019;267:1-38.

Floridi L, Cowls J. A unified framework of five principles for AI in society. *Machine learning and the city: Applications in architecture and urban design*. 2022:535-45.

O'Neill T, McNeese N, Barron A, Schelble B. Human–autonomy teaming: A review and analysis of the empirical literature. *Hum Factors*. 2022;64(5):904-38.

Reda HT, Anwar A, Mahmood A. Comprehensive survey and taxonomies of false data injection attacks in smart grids: Attack models, targets, and impacts. *Renew Sustain Energy Rev*. 2022;163:112423.

Ding J, Qammar A, Zhang Z, Karim A, Ning H. Cyber threats to smart grids: Review, taxonomy, potential solutions, and future directions. *Energies*. 2022;15(18):6799.

Xu C, Liao Z, Li C, Zhou X, Xie R. Review on interpretable machine learning in smart grid. *Energies*. 2022;15(12):4427.

Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ. XAI—Explainable artificial intelligence. *Sci Robot*. 2019;4(37):eaay7120.

Ucar F. A comprehensive analysis of smart grid stability prediction along with explainable artificial intelligence. *Symmetry*. 2023;15(2):289.

Yayla A, Haghnegahdar L, Dincelli E. Explainable artificial intelligence for smart grid intrusion detection systems. *IT Prof*. 2022;24(5):18-24.

Johnson CJ, Demir M, McNeese NJ, Gorman JC, Wolff AT, Cooke NJ. The impact of training on human–autonomy team communications and trust calibration. *Hum Factors*. 2023;65(7):1554-70.

McNeese NJ, Demir M, Cooke NJ, Myers C. Teaming with a synthetic teammate: Insights into human-autonomy teaming. *Hum Factors*. 2018;60(2):262-73.

Aoufi S, Derhab A, Guerroumi M. Survey of false data injection in smart power grid: Attacks, countermeasures and challenges. *J Inf Secur Appl*. 2020;54:102518.

Asmat MN, Khan SU, Hussain S. Uncertainty handling in cyber–physical systems: State-of-the-art approaches, tools, causes, and future directions. *J Softw Evol Process*. 2023;35(7):e2428.

Unsal DB, Ustun TS, Hussain SS, Onen A. Enhancing cybersecurity in smart grids: False data injection and its mitigation. *Energies*. 2021;14(9):2657.

Du D, Zhu M, Li X, Fei M, Bu S, Wu L, et al. A review on cybersecurity analysis, attack detection, and attack defense methods in cyber-physical power systems. *J Modern Power Syst Clean Energy*. 2022;11(3):727-43.

Diaba SY, Shafie-Khah M, Elmusrati M. Cyber-physical attack and the future energy systems: A review. *Energy Rep*. 2024;12:2914-32.

Cifci A. Interpretable prediction of a decentralized smart grid based on machine learning and explainable artificial intelligence. *IEEE access*. 2025;13:36285-305.