

Public Trust in Automated Decision Systems

Ethan Ward - Corresponding author

Abstract

This interdisciplinary review examines public trust in automated decision systems. The organizing question is how competence, procedural fairness, transparency, recourse, and institutional legitimacy shape warranted trust. Ten related scholarly sources are synthesized through a decision-centered framework spanning problem definition, mechanism, measurement, evaluation, implementation, and governance. The review does not invent experiments, pooled estimates, or unreported quantitative results. It instead evaluates the strength and transferability of the available evidence, with particular attention to treating acceptance or adoption as equivalent to justified trust. The resulting framework links technical or empirical performance to explicit use conditions and identifies tests that should precede wider adoption in public-sector and socially consequential automation.

Keywords: public trust, automated decisions, procedural fairness, accountability, institutions

Synthesis lens	Principal question
Mechanism	the chain connecting evidence inputs, public trust in automated decision systems, intermediate outputs, and downstream decisions
Evaluation	construct validity, comparative performance, uncertainty, transfer across settings, and decision utility
Primary risk	treating acceptance or adoption as equivalent to justified trust

1. Introduction

Research on public trust in automated decision systems sits at the boundary between a promising component and a consequential decision. The core question is how competence, procedural fairness, transparency, recourse, and institutional legitimacy shape warranted trust. Answering it requires more than assembling favorable results. It requires a chain of evidence that makes the problem boundary, mechanism, measurement, comparator, uncertainty, and accountable decision visible to the reader.

This article presents a structured narrative review of ten related publications. Sources were selected for topical relevance, traceable publication metadata, and complementary coverage of technical, empirical, and governance questions. No meta-analytic pooling is attempted because the included studies differ in designs, populations, system boundaries, and outcomes. The purpose is decision-oriented synthesis: to identify what each source can support, where transfer remains uncertain, and which evidence would justify the next stage of use.

The central risk is treating acceptance or adoption as equivalent to justified trust. A top-line metric or broad policy label can appear decisive while concealing the conditions that produced it. Accordingly, the synthesis separates observation from interpretation and implementation from validation. This separation is especially important when the same system creates benefits for one user or institution while transferring cost, risk, or administrative burden to another.

2. Evidence Base and Problem Boundary

A defensible review begins by defining the unit of analysis. For the present topic, the relevant boundary includes inputs, institutional or technical context, the mechanism producing an output, the person or system acting on that output, and the downstream outcome. Evidence falls outside the boundary when it measures only an intermediate proxy yet is used to claim an end-to-end benefit.

Araujo et al. (2020) addresses "In AI we trust? Perceptions about automated decision-making by artificial intelligence." In this synthesis, that work is used as a source on problem definition within public trust in automated decision systems. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for public-sector and socially consequential automation. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Kuziemski and Misuraca (2020) addresses "AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings." In this synthesis, that work is used as a source on conceptual boundary within public trust in automated decision systems. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for public-sector and socially consequential automation. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Kyriakidis et al. (2015) addresses "Public opinion on automated driving: Results of an international questionnaire among 5000 respondents." In this synthesis, that work is used as a source on measurement within public trust in automated decision systems. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for public-sector and socially consequential automation. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Together, these works provide complementary entry points, but their conclusions should not be flattened into a single ranking. Differences in data generation, setting, temporal horizon, and outcome definition may be substantively meaningful. A review should therefore preserve those differences and state where analogy, rather than direct replication, connects one domain to another.

3. Mechanism and System Design

The operative mechanism is the chain connecting evidence inputs, public trust in automated decision systems, intermediate outputs, and downstream decisions. This formulation discourages component-only reasoning. A model, platform, policy, assay, or infrastructure service acquires practical meaning only when its interfaces and use conditions are specified. The evidence chain should describe what information is available, what transformation occurs, which assumptions make that transformation credible, and who is authorized to act.

Winfield and Jirotko (2018) addresses "Ethical governance is essential to building trust in robotics and artificial intelligence systems." In this synthesis, that work is used as a source on system mechanism within public trust in automated decision systems. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for public-sector and socially consequential automation. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Licht and Licht (2020) addresses "Artificial intelligence, transparency, and public decision-making." In this synthesis, that work is used as a source on representation choice within public trust in automated decision systems. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for public-sector and socially consequential automation. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Lăzăroiu et al. (2020) addresses "Consumers' Decision-Making Process on Social Commerce Platforms: Online Trust, Perceived Risk, and Purchase Intentions." In this synthesis, that work is used as a source on failure analysis within public trust in automated decision systems. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for public-sector and socially consequential automation. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

These studies also show why modular claims require interface tests. A component may perform well while the complete pathway fails because inputs drift, users interpret outputs differently, recovery semantics are ambiguous, or institutional incentives change. Design documentation should therefore include not only the intended pathway but also foreseeable deviations, degraded modes, and conditions under which the system should stop or defer.

4. Evaluation and Uncertainty

Evaluation should center on construct validity, comparative performance, uncertainty, transfer across settings, and decision utility. Average performance remains useful, but it should be accompanied by distributions, error categories, relevant subgroups or operating regimes, and uncertainty at the point where a decision changes. Comparators must isolate the value of the proposed addition rather than compare a mature intervention with an intentionally weak baseline.

Engin and Treleaven (2018) addresses "Algorithmic Government: Automating Public Services and Supporting Civil Servants in using Data Science Technologies." In this synthesis, that work is used as a

source on comparative evaluation within public trust in automated decision systems. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for public-sector and socially consequential automation. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Ryan (2020) addresses "In AI We Trust: Ethics, Artificial Intelligence, and Reliability." In this synthesis, that work is used as a source on uncertainty within public trust in automated decision systems. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for public-sector and socially consequential automation. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

External validation should introduce meaningful shifts in setting, time, population, workload, market structure, or environmental conditions. A nominally independent sample can still be too similar to test transfer. Conversely, a failed transfer is scientifically informative when the changed conditions are measured and the mechanism of failure is examined rather than hidden in an aggregate score.

5. Translation and Governance

Translation to public-sector and socially consequential automation depends on more than technical readiness. Decision rights, documentation, maintenance, monitoring, appeal, and recovery are part of the intervention. The governance requirement is traceable evidence, named responsibility, version control, monitoring, appeal, and explicit revalidation. Each control should be connected to a named failure mode and should identify an owner, evidence source, review interval, and escalation path.

Azevedo-Sa et al. (2020) addresses "Real-Time Estimation of Drivers' Trust in Automated Driving Systems." In this synthesis, that work is used as a source on implementation within public trust in automated decision systems. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for public-sector and socially consequential automation. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Robinson (2020) addresses "Trust, transparency, and openness: How inclusion of cultural values shapes Nordic national public policy strategies for artificial intelligence (AI)." In this synthesis, that work is used as a source on governance within public trust in automated decision systems. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for public-sector and socially consequential automation. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

A credible implementation plan also defines sunset and revalidation conditions. New data, software updates, institutional restructuring, population change, or shifts in incentives can invalidate previously reasonable assumptions. Monitoring should therefore test the validity of the evidence chain, not merely whether a service remains online or a headline metric remains stable.

6. Research Agenda

Future studies should preregister or otherwise predefine the intended decision, principal outcomes, exclusions, and analysis plan when feasible. They should report missingness, sensitivity analyses, negative or null findings, and the cost of errors to differently situated stakeholders. Reproducible artifacts should include sufficient metadata to reconstruct data provenance, model or analytical versions, parameter choices, and the sequence from evidence to conclusion.

Cross-disciplinary work needs a shared object of explanation rather than a collection of adjacent vocabularies. For public trust in automated decision systems, that shared object is the complete decision pathway. Technical, clinical, social, environmental, or economic expertise should each change the design or interpretation of the study. When evidence remains incomplete, the useful output is a bounded hypothesis, a transparent uncertainty statement, or a request for additional measurement.

7. Conclusion

The literature supports a disciplined approach to public trust in automated decision systems: define the decision, preserve system boundaries, test consequential failure modes, connect uncertainty to action, and assign accountable control. The strongest conclusion is not that one method is universally superior. It is that credible use in public-sector and socially consequential automation requires an auditable link from source evidence to design choice, evaluation result, and governed decision.

Declarations

Ethics approval: Not applicable. This article synthesizes previously published literature and reports no new research involving humans, animals, human tissue, or identifiable sensitive data.

Consent to participate and consent for publication: Not applicable.

Funding: No external funding is reported for this commissioned synthesis.

Competing interests: The authoring group declares no competing interests for this commissioned synthesis.

AI-assisted writing: Generative AI supported drafting, source organization, and language editing. Accountable authors and editors must verify every claim, citation, and disclosure before publication.

Data, code, and materials availability: No new dataset or code was generated. All evidence is identified in the reference list.

Licence: This manuscript is prepared for publication under the Creative Commons Attribution 4.0 International licence (CC BY 4.0).

References

- Araujo, T., Helberger, N., Kruijemeier, S., & Vreese, C. H. D. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & Society*, 35(3), 611-623. <https://doi.org/10.1007/s00146-019-00931-w>
- Azevedo-Sa, H., Jayaraman, S. K., Esterwood, C. T., Yang, X. J., Robert, L. P., & Tilbury, D. M. (2020). Real-Time Estimation of Drivers' Trust in Automated Driving Systems. *International Journal of Social Robotics*, 13(8), 1911-1927. <https://doi.org/10.1007/s12369-020-00694-1>
- Engin, Z., & Treleaven, P. (2018). Algorithmic Government: Automating Public Services and Supporting Civil Servants in using Data Science Technologies. *The Computer Journal*, 62(3), 448-460. <https://doi.org/10.1093/comjnl/bxy082>
- Kuziemski, M., & Misuraca, G. (2020). AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. *Telecommunications Policy*, 44(6), 101976. <https://doi.org/10.1016/j.telpol.2020.101976>

- Kyriakidis, M., Happee, R., & Winter, J. C. F. D. (2015). Public opinion on automated driving: Results of an international questionnaire among 5000 respondents. *Transportation Research Part F Traffic Psychology and Behaviour*, 32, 127-140. <https://doi.org/10.1016/j.trf.2015.04.014>
- Licht, K. D. F., & Licht, J. D. F. (2020). Artificial intelligence, transparency, and public decision-making. *AI & Society*, 35(4), 917-926. <https://doi.org/10.1007/s00146-020-00960-w>
- Lăzăroiu, G., Neguriță, O., Grecu, I., Grecu, G., & Mitran, P. C. (2020). Consumers' Decision-Making Process on Social Commerce Platforms: Online Trust, Perceived Risk, and Purchase Intentions. *Frontiers in Psychology*, 11, 890. <https://doi.org/10.3389/fpsyg.2020.00890>
- Robinson, S. C. (2020). Trust, transparency, and openness: How inclusion of cultural values shapes Nordic national public policy strategies for artificial intelligence (AI). *Technology in Society*, 63, 101421. <https://doi.org/10.1016/j.techsoc.2020.101421>
- Ryan, M. (2020). In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Science and Engineering Ethics*, 26(5), 2749-2767. <https://doi.org/10.1007/s11948-020-00228-y>
- Winfield, A. F. T., & Jirotko, M. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A Mathematical Physical and Engineering Sciences*, 376(2133), 20180085. <https://doi.org/10.1098/rsta.2018.0085>