frontiers in Political Science

TYPE Original Research
PUBLISHED 23 November 2026
DOI 10.3389/fpos.2026.1904617

Check for updates

OPEN ACCESS

EDITED BY
Fatemeh Ahmadi Zeleti,
Insight Centre for Data Analytics, Ireland

REVIEWED BY
Shima Taheri,
Islamic Azad University, Isfahan, Iran
Avram Sorin,
University of Craiova, Romania

*CORRESPONDENCE
Koen Smit
koen.smit@hu.nl

RECEIVED 09 June 2026
REVISED 04 September 2026
ACCEPTED 19 October 2026
PUBLISHED 23 November 2026

CITATION
Smit K, Peters R, Leewis S, van Meerten J, Kok F and Grouls R (2026) When is generative AI ready for spatial planning? A value-sensitive, mixed-methods evaluation of LLM-based wind-energy mapping demonstrator.
Front. Polit. Sci. 8:1904617.
doi: 10.3389/fpos.2026.1904617

COPYRIGHT
© 2026 Smit, Peters, Leewis, van Meerten, Kok and Grouls. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

When is generative AI ready for spatial planning? A value-sensitive, mixed-methods evaluation of LLM-based wind-energy mapping demonstrator

Koen Smit[1]*, Rob Peters[2], Sam Leewis[1], John van Meerten[1], Fabian Kok[1] and Raoul Grouls[2]

[1]HU University of Applied Sciences Utrecht, Research group Digital Ethics, Utrecht, Netherlands, [2]Province of Utrecht, Utrecht, Netherlands

Introduction: Generative AI is increasingly proposed to support spatial planning, yet empirical evidence on how such systems perform in realistic planning workflows and how professionals evaluate their potential use remains limited. This study reports an exploratory mixed-method case evaluation of a generative AI–based demonstrator that translates legal and policy text together with geospatial inputs into a machine-readable spatial artefact, illustrated through a wind-energy opportunity-mapping case for the Province of Utrecht.
Methods: The evaluation combined a qualitative expert comparison with an existing professionally produced reference map, two purposively composed expert focus groups, descriptive Technology Acceptance Model (TAM) questionnaires, a public-values ranking, and reflexive thematic analysis. Value Sensitive Design (VSD) was used as an empirical and interpretive lens to identify public and professional values relevant to further development of the demonstrator and underlying workflow.
Results: The reference comparison showed that the demonstrator could generate a structurally recognizable spatial output, but also revealed crucial discrepancies, including missing buffers, insufficient differentiation of constraint categories, and weak grounding of some policy-based constraints. Questionnaire and focus-group findings indicate that participants recognized potential usefulness, particularly for rapid exploration and synthesis, while simultaneously emphasizing the need for validation, traceability, controllability, and human oversight.
Discussion: Given the small expert samples and exploratory design, these findings should not be interpreted as evidence of professional technology adoption. The evaluated demonstrator should instead be understood as an exploratory proof of concept whose outputs are not yet sufficiently reliable for authoritative spatial-planning use. Quantitative spatial validation, reproducibility testing, and stronger integration of deterministic validation mechanisms remain necessary next steps.

KEYWORDS
artificial intelligence, digital twins, generative AI, large language models, planning support systems, spatial planning, technology acceptance, value-sensitive design

10.3389/fpos.2026.1904617

1 Introduction

Recent advances in generative AI are opening new possibilities for urban planning. Large language models (LLMs) and multimodal generative networks (e.g., GPT, Claude, Gemini, Llama, Mistral, and diffusion models such as DALL-e and Sora) can now process complex inputs (text, spatial data, imagery) and generate design alternatives or analyses. Surveying this trend, Pan et al. (2025) report that LLMs are transforming urban planning from passive, tool-based processes into active human–AI collaborations: they enable "natural language–driven design generation" and multi-agent stakeholder simulations. In practical terms, planners can use conversational AI to sketch layouts or simulate scenarios on-the-fly, making the design process more interactive and dynamic.

Spatial planning is inherently a multi-stakeholder, value-laden "wicked problem" requiring trade-offs among economic, social, and environmental goals, see, e.g., Figure 1. Planners must balance competing objectives (e.g., housing vs. green space) as well as public values like equity, participation, and environmental sustainability. For example, in the Netherlands, where building space is limited, multiple societal issues also heavily steer decision-making, such as the housing shortage, migration, nitrogen deposition, infrastructure renewal and maintenance deficit, electricity grid congestion, and physical and social safety. Brama et al. (2024) note that urban design tasks often involve ambiguous requirements and conflicting criteria, aligning with the "wicked" complexity described by Rittel and Webber (1973). Traditional decision-support uses multi-criteria analysis to compare options, but this can be slow and may overlook qualitative and geometric concerns. Therefore, there is a need for decision-support tools that can handle large, heterogeneous datasets and incorporate planning values in an integrated way.

In practice, spatial planning decisions are often developed through sectoral analyses conducted by different disciplinary domains and urban policy themes, such as water management, infrastructure, environmental protection, and public safety. The results of these analyses are typically documented in static planning reports, policy documents, legislative documents and technical studies, which are subsequently aggregated by policy officers, portfolio holders, and political decision-makers when evaluating spatial alternatives or planning scenarios. While this process allows domain experts to contribute specialized knowledge and support deliberative governance, it also introduces several limitations. Knowledge remains fragmented across documents and disciplines, cross-domain trade-offs are difficult to systematically evaluate, and decision-making relies heavily on manual interpretation of heterogeneous information sources. Such characteristics have long been recognized as challenges for integrated spatial planning and have motivated the development of digital planning support systems aimed at improving data integration and scenario exploration (Batty, 1995; Geertman and Stillwell, 2004; Klosterman, 1997).

Generative AI provides new possibilities for supporting some of these information-intensive planning tasks. Recent work suggests that LLM-based systems can assist with activities such as information synthesis, scenario generation, data analysis, and stakeholder communication (Jiang et al., 2025; Pan et al., 2025). However, these capabilities should not be understood as resolving the underlying complexity of spatial planning. Fragmentation in



FIGURE 1
Spatial planning wicked problem space (De Lathouwer et al., 2024).

planning is not merely an information-processing problem. It also reflects institutional responsibilities, legal interpretation, political deliberation, professional judgement, and competing public values (Lartey and Law, 2025; Rittel and Webber, 1973). Generative AI may therefore support specific analytical and exploratory tasks within planning processes, but it does not remove the need for domain expertise or normative and institutional decision-making. This distinction is particularly important in legally constrained planning contexts, where a technically plausible output may still be substantively or procedurally incorrect (Wang et al., 2026).

Despite this potential, significant uncertainty remains about how generative AI performs when embedded in integrated spatial-planning workflows rather than evaluated on isolated language or geospatial tasks (Li et al., 2025; Van de Weghe et al., 2025; Zhang et al., 2025). This is very relevant in legally and institutionally constrained settings, where errors in constraint classification, spatial buffering, or policy interpretation may undermine the reliability and institutional acceptability of generated planning artefacts (Wang et al., 2026). At the same time, professional evaluation of such systems cannot be reduced to technical performance alone, because their acceptability is also shaped by transparency, controllability, accountability, trust, and other public and professional values (Chen et al., 2023; Wanner et al., 2022; Zuiderwijk et al., 2021).

Taking into account these challenges, this study examines an exploratory generative AI–based demonstrator for translating policy, legal, and geospatial information into a wind-energy opportunity map. Our research question therefore is as follows: To what extent can a generative AI–based workflow support exploratory spatial-planning tasks, and which technical, professional, and public-value conditions shape experts' assessment of its potential use?

This paper makes three bounded contributions. First, it provides an empirically grounded case evaluation of an integrated generative AI spatial-planning workflow in a governmental setting, including a qualitative comparison between the generated artefact and an existing professional reference map. Second, it provides case-specific evidence on how spatial-planning, GIS, policy, and AI experts assess the potential usefulness and limitations of such a workflow. Third, it uses TAM and VSD for complementary analytical purposes: TAM provides descriptive information on individual perceptions of usefulness and ease of use, while VSD supports the elicitation and interpretation of normative and institutional requirements. The study therefore does not propose a new general theory of AI adoption but provides contextual empirical evidence concerning the conditions under which this type of planning-support technology may warrant further development and validation. The remainder of this paper is structured as follows. Section 2 reviews the literature on planning support systems, digital twins, generative AI–enabled geospatial decision support, technology acceptance, and public values. Section 3 describes the research design, demonstrator, expert focus groups, measures, and analytical procedures. Section 4 presents the empirical results, including the qualitative comparison with the professional reference map and the functional, governance, and technical expert evaluations. Section 5 discusses the findings and study limitations. Section 6 concludes the paper, and Section 7 outlines the future research agenda.

2 Background and related work

2.1 Multimodal data in spatial planning

Spatial planning is fundamentally a multimodal knowledge domain. Planners must routinely combine and interpret heterogeneous information sources, including geospatial vector and raster datasets, remote-sensing imagery, maps and 2D/3D city models, legal and policy texts, zoning ordinances, environmental regulations, and formalized indicators or rule systems. This makes planning qualitatively different from domains in which a single modality dominates. Recent reviews show that the role of generative AI in geospatial and planning contexts is evolving from general-purpose text generation toward domain-specific support for spatial analysis, workflow automation, urban scenario generation, and multimodal reasoning (Li et al., 2025; Wang et al., 2024). This multimodal character is not merely descriptive. It has direct implications for AI design. A development proposal, for example, may require simultaneous interpretation of parcel geometry, land-use layers, transport accessibility, environmental constraints, statutory text, and visual design alternatives. Reviews at the GIScience–LLM interface therefore emphasize that current models still struggle with spatial semantics, topology, scale, uncertainty, and multimodal grounding, even when they perform well on generic language tasks (Wang et al., 2024; Van de Weghe et al., 2025). Similarly, work on urban digital twins suggests that generative models are currently strongest in content generation, such as scenario, design, or synthetic data production, but much weaker in regulatory compliance, context-sensitive reasoning, and transparent justification (Batty, 2024; Xu et al., 2024).

A recurring finding across this literature is that linguistic fluency does not equal spatial competence. This distinction is directly related to the problem of hallucination in generative AI. When models rely primarily on linguistic pattern recognition without sufficient grounding in spatial data or formal rules, they tend to generate outputs that appear coherent but do not accurately reflect spatial relationships or regulatory constraints. In spatial planning contexts, this results in outputs that are superficially plausible yet incorrect. This limitation is not incidental but structural. It reflects the architecture of neural AI, where outputs are generated probabilistically rather than through deterministic, rule-based reasoning. As a result, standalone LLMs are not sufficiently reliable for tasks that require precise spatial reasoning, consistency, and legally grounded interpretation. Empirical studies confirm that such models perform poorly on GIS-relevant spatial cognition tasks, especially those involving route and survey knowledge, geometric constraints, or multi-step reasoning. Performance improves substantially when LLMs are combined with deterministic GIS functions, external map data, or structured retrieval mechanisms (Manvi et al., 2024; Yang et al., 2025). This should not be interpreted merely as an empirical improvement, but as evidence that neural and symbolic approaches fulfill complementary roles. Neural models support flexible multimodal interpretation of source documentation, while symbolic systems provide the formal structure, consistency, and verifiability required for spatial analysis and decision-making. The central research challenge is therefore not whether LLMs can summarize planning documents, but how to design integrated, neuro-symbolic workflows that reliably connect policy text, spatial data, and formal decision logic within transparent and validated planning processes.

2.2 Planning support systems and digital twins in spatial planning

AI-enabled spatial planning should also be understood within a longer trajectory of computational change in the discipline. A recent 30-year review identifies a progression from rational-comprehensive approaches toward increasingly data-driven and algorithmic forms of planning and, more recently, toward a hybrid socio-technical phase in which AI mediates between technical infrastructures, professional knowledge, and participatory processes (Shabani et al., 2025). Generative AI therefore represents a recent development within a much longer evolution of computational planning support rather than an entirely new departure. The most important historical foundation for AI-enabled planning tools is the literature on Planning Support Systems (PSS). Classic PSS research defined the field as a family of geo-information-based methods and tools that support planning analysis, communication, and decision-making, while repeatedly documenting a persistent implementation gap between technically sophisticated tools and everyday planning practice (Geertman and Stillwell, 2004, 2020). Over time, this literature evolved from a narrow focus on computational systems toward a broader planning support science that pays greater attention to users, institutional context, mediation, and the practical embedding of tools in planning processes (Geertman and Stillwell, 2020). This shift is crucial for understanding the prospects of generative AI in planning: technical capability alone has never been sufficient for uptake in practice. Later PSS work reinforced this point by showing that the usefulness of planning technologies often depends less on predictive or analytic sophistication than on whether they improve communication, collaboration, learning, and deliberation. Brömmelstroet and Schrijnen (2010) argued for a move from planning support systems toward mediated planning support, emphasizing structured dialogue between developers and users. Pelzer et al. (2014) similarly found that practitioners often value PSS for their contribution to shared understanding and collaborative decision-making rather than solely for better-informed technical outputs. These findings remain directly relevant to generative AI: planning tools will likely be judged not just on whether they automate tasks, but on whether they improve scenario comparison, explanation of trade-offs, and communication among planners, experts, decision-makers, and stakeholders. This is in line with our prototype and research, which is not aimed to replace spatial planners, but to support their spatial planning processes.

Urban digital twins can be understood as a next-generation extension of the PSS development cycle. Compared with traditional PSS, digital twins promise tighter coupling between the city and its computational representation through (real-time) data integration, simulation, visualization, scenario modelling, and iterative feedback. Batty (2024) positions digital twins as an important new planning infrastructure, but also cautions that they are not neutral mirrors of reality. They are selective computational representations shaped by modelling assumptions and human choices. Governance-oriented digital twin research similarly emphasizes that maturity depends not only on technical performance, but also on interoperability, participation, transparency, inclusivity, and explicit treatment of human judgment (Haraguchi et al., 2024; Therias and Rafiee, 2023). In other words, digital twins extend PSS technologically, but they do not resolve the socio-technical and legislative problems that have always limited planning-support adoption. For this paper, the key insight is that digital twins should not be framed as replacing PSS, but as expanding the planning-support paradigm into data-rich, and potentially AI-enabled environments. Yet the same long-standing challenges remain central: implementation in practice, trust, institutional fit, data integration, and meaningful public trust. These issues are particularly important if generative AI is to be embedded within digital twins as a decision-support layer rather than merely as an interface or visualization add-on.

2.3 Generative AI and LLM-based geospatial decision support

A rapidly growing body of work now studies LLMs not simply as text interfaces, but as the reasoning layer in AI-assisted geospatial decision support. The most ambitious version of this agenda is often described as autonomous GIS, in which an LLM parses a user goal, retrieves relevant spatial data, generates a geoprocessing workflow, executes tools, and returns maps, code, or narrative interpretations (Li et al., 2025). Related agenda-setting work argues that the integration of GIS-science and LLMs may constitute a major new interdisciplinary frontier, but also stresses that current systems remain early-stage prototypes rather than mature decision-support infrastructures (Van de Weghe et al., 2025).

Empirical studies reinforce both the promise and the limitations of this direction. GeoLLM demonstrates that LLMs contain latent geospatial knowledge that can be extracted for prediction tasks when grounded in auxiliary map data such as OpenStreetMap (Manvi et al., 2024). Tool-enabled systems such as AutoGEEval and ShapefileGPT show that LLMs can support geospatial code generation, vector data processing, and workflow automation, particularly when embedded in agentic architectures or constrained by formal APIs (Lin et al., 2025; Wu et al., 2025). GeoAnalystBench, meanwhile, provides much-needed benchmarking evidence, showing that leading proprietary models can produce valid spatial workflows for many tasks, but still struggle on tasks requiring deeper spatial reasoning such as optimal site selection or spatial relationship detection (Zhang et al., 2025). More generally, evaluation studies indicate that the best current systems perform most reliably when LLMs are combined with external tools, structured task decomposition, and explicit validation mechanisms rather than being asked to reason by the model alone. This pattern can be understood as a response to hallucination risks. By constraining model outputs through deterministic tools, structured workflows, and validation steps, developers aim to reduce the likelihood of plausible but incorrect spatial or policy interpretations, which is in line with, e.g., Zhang et al. (2025).

Despite these advances, the literature remains highly fragmented across isolated tasks. Existing studies typically demonstrate one capability at a time: map generation, code generation, spatial question answering, shapefile operations, or geospatial prediction. Much less work addresses the integrated workflows that matter in spatial planning, where decision support must connect policy retrieval, legal interpretation, geospatial analysis, indicator calculation, scenario evaluation, and explainable and transparent recommendation. This gap is particularly consequential because planning decisions are not only spatially complex but also institutionally constrained. Adjacent legal-knowledge-graph research suggests one promising bridge: coupling LLMs with structured knowledge representations to support

more reliable extraction and reasoning over regulatory text (Li et al., 2024). For spatial planning, the open challenge is therefore not only autonomous GIS, but policy-aware, explainable, and institutionally grounded geospatial decision support.

2.4 Acceptance, trust, and public values in government AI systems

Research on the adoption of AI in government builds on the broader technology-acceptance tradition, especially the Technology Acceptance Model (TAM) line, which identifies perceived usefulness and perceived ease of use as core determinants of uptake (Davis, 1989). Yet studies of public-sector AI consistently show that adoption depends on more than these classical variables. In public administration, AI uptake is shaped by organizational readiness, leadership support, skills, data quality, and institutional fit, but also by a distinctive set of public values, including transparency, explainability, accountability, procedural fairness, controllability, and legitimacy (Chen et al., 2023; Neumann et al., 2024). These values are not peripheral ethical extras, they are central to whether AI-assisted decision tools are regarded as acceptable in domains such as spatial planning, where decisions are politically contestable and affect the public interest. Trust is therefore not merely a psychological afterthought but a constitutive condition of public-sector AI acceptance. Reviews of trust in AI distinguish technical trustworthiness factors such as safety, robustness, and performance from non-technical factors such as fairness, dignity, explainability, and human control (Afroogh et al., 2024). Studies of intelligent decision-support systems show that transparency often affects acceptance indirectly, by shaping trust and perceived usefulness, rather than simply increasing adoption by default (Wanner et al., 2022). In government contexts, these issues are intensified because AI systems must remain reviewable, contestable, and accountable to democratic institutions and legal norms. Recent governance research therefore argues that good public-sector AI must support not only efficiency and effectiveness, but also procedural justice, social justice, responsiveness, resilience, and counterbalance (Mišić et al., 2025). For AI-enabled planning systems, this means that acceptance cannot be evaluated solely in terms of user satisfaction or task efficiency; it must also be assessed in relation to accountability, contestability, and alignment with public values.

This is where Value Sensitive Design (VSD) becomes especially relevant. Developed by Friedman and colleagues, VSD treats technologies as inherently value-laden and proposes a tripartite methodology of conceptual, empirical, and technical investigations to identify stakeholders, surface value tensions, and translate values into design requirements (Friedman et al., 2002, 2017; Friedman and Hendry, 2019). In contrast to high-level ethical principle lists, VSD offers a more design-oriented and socio-technical framework for embedding values into requirements, interfaces, data practices, evaluation protocols, and governance arrangements. Recent reviews of value-sensitive AI argue that existing AI ethics approaches often remain too abstract for practical design, whereas VSD helps make values such as transparency, fairness, accountability, privacy, and human agency operational throughout the lifecycle of a system (Sadek et al., 2024; Simon et al., 2020). In public-sector settings, VSD aligns closely with public-values research by emphasizing legislative accuracy, stakeholder inclusion, anticipation of (in)direct harms, and explicit handling of value conflicts.

For the this study, TAM and VSD are therefore used for distinct but complementary analytical purposes rather than combined into a single technology-adoption model. TAM is used narrowly to describe participating experts' individual perceptions of usefulness and ease of use (Davis, 1989; Venkatesh and Bala, 2008). However, by itself it does not capture institutional legitimacy, legal acceptability, public values, or governance readiness. VSD addresses a different analytical level by providing a socio-technical framework for identifying and interpreting values, stakeholder concerns, and design requirements (Friedman et al., 2017; Friedman and Hendry, 2019). The mixed-method design brings these perspectives together only at the level of interpretation as descriptive acceptance perceptions are considered alongside qualitative evidence concerning the normative, institutional, and technical conditions participants associated with acceptable use.

Recent research on AI adoption in urban governance similarly indicates that practical uptake is conditioned not only by technological capability but also by institutional knowledge structures, organizational readiness, governance arrangements, and social and ethical considerations (Lartey and Law, 2025, 2026). This literature suggests that the transition from technical possibility to institutional use is itself a socio-technical governance challenge. The present study adds to this literature by examining these conditions from the perspective of experts evaluating a concrete generative-AI planning demonstrator.

3 Research method

3.1 Overall research design

This study employs an exploratory mixed-method case-study design combining qualitative expert appraisal, reflexive thematic analysis of expert deliberations, and a descriptive measurement of perceived usefulness and ease of use. The design is exploratory because the study aims to examine a developing technology in context rather than test pre-specified causal hypotheses or estimate population-level adoption effects (Creswell and Plano Clark, 2018). The study uses two complementary expert perspectives rather than treating FG1 and FG2 as statistically comparable samples or as a formal replication design. FG1 was oriented toward functional use, planning context, governance, and public values, whereas FG2 emphasized technical feasibility, validation requirements, architecture, and spatial correctness. Questionnaire data are therefore analyzed descriptively within each FG, while the qualitative material provides contextual insight into how participating experts understood the demonstrator's potential, limitations, and requirements for further development. This mixed-method interpretation is exploratory and does not test population-level differences or causal hypotheses (Creswell and Plano Clark, 2018).

The design operationalizes two principles central to rigorous sociotechnical evaluation. First, that technology maturity judgements must be evidence-anchored, linked to a concrete demonstration environment, and elicited through structured expert deliberation (Lavin et al., 2022; Mankins, 1995). Second, that acceptance of AI-enabled decision-support tools in professional and governmental contexts is jointly determined by demonstrated capability, perceived utility, and trust-related constraints that condition appropriate reliance, dimensions not reducible to a single measurement instrument (Davis, 1989; Wanner et al., 2022; Zuiderwijk et al., 2021). The study design is summarized in Table 1.

TABLE 1 Study design overview.

Dimension	FG1: functional focus	FG2: technical focus
Date	October 2025	November 2025
Participants	20	12
Primary emphasis	Policy-cycle alignment; governance risks; domain utility	Apparent geometric alignment; reproducibility requirements; validation feasibility
Participant profiles	Provincial GIS, Omgevingswet policy, digital-twin strategy, AI R&D, National representatives, ministry of spatial affairs, Geonovum formal standardization agency	GIS specialists, AI practitioners, applied research, safety regions
TAM role	Descriptive TAM assessment	Complementary descriptive TAM assessment
Emphasis	Functional and governance readiness	Technical and validation readiness

3.2 Case study organization, materials and prototype (demonstrator)

The study was conducted within the Province of Utrecht, a regional government authority in the Netherlands responsible for coordinating spatial, environmental, and infrastructural policy at the supra-municipal level. The political-administrative system is supported by a civil-service organization and an executive management layer, which together form the institutional setting in which spatial planning decisions are prepared, evaluated, and implemented. Within this context, provincial responsibilities include the development and coordination of integrated spatial policies under the Dutch Omgevingswet, requiring the combination of policy interpretation, identification and classification of legal constraints, and geospatial analysis into concrete planning outputs. The study was conducted primarily with the Province of Utrecht's geo-information (GIS) department, which is responsible for managing and publishing spatially referenced information used across provincial themes and planning tasks. The province operationalizes this through (1) a digital provincial ATLAS that enables interactive consultation and layer-combination across provincial themes and provides downloads of underlying datasets where they are publicly available, and (2) the Geo-Point Utrecht portal that bundles geo-apps and datasets and explicitly positions provincial geodata for reuse via downloadable formats and geoservices. Positioning the case within this GIS function clarifies why the demonstrator's outputs were evaluated against an existing professionally produced opportunity map. The work is situated in an organizational context whose routine practice includes translating policy/legal constraints into spatial layers and map products and distributing them through established provincial geo-information channels.

The primary object of evaluation in both FGs was a generative AI–based spatial decision-support workflow designed to produce a geospatial opportunity map for wind-energy development (hereafter: the demonstrator). The demonstrator combines (1) policy and legal text excerpts describing hard, complex, and soft constraints with (2) geospatial source layers and uses a large language model (LLM) to generate a machine-readable geospatial artefact. Throughout the case evaluation, "hard constraints" refer to exclusionary legal or policy constraints that preclude development at a location, "complex constraints" refer to conditional constraints that require additional case-specific assessment, and "soft constraints" refer to lower-severity considerations generated by the demonstrator that may affect suitability without constituting a direct exclusion. The professional reference case did not contain a directly equivalent soft-constraint category, which limits direct comparison for this category. The demonstrator encodes constraint categories as spatial geometries (e.g., exclusion zones and buffered constraint areas) and delineates residual opportunity space. For presentation and review, the generated output was visualized in QGIS and subjected to an expert-based comparison with a reference opportunity map previously produced in routine practice by GIS professionals within the Province of Utrecht, see Figure 2.

The task within the case organization was selected because it represents a common and decision-relevant planning-support task in which legal interpretation, policy reasoning, and geospatial processing must be integrated into a concrete planning product (Geertman and Stillwell, 2020; Li et al., 2025). This makes it suitable for evaluating whether a generative AI workflow can contribute beyond generic text assistance by producing an output that resembles a recognizable GIS artefact used in professional practice. The qualitative reference comparison enabled experts to identify apparent similarities and discrepancies between the generated artefact and the professionally produced opportunity map and to discuss the conditions under which such outputs could potentially be validated and used. It was not designed as a quantitative spatial-accuracy benchmark.

The demonstrator evaluated in this study is an exploratory, expert-facing research prototype. Its intended role is to support rapid exploration, synthesis, and early-stage identification of potentially relevant spatial constraints. It is not designed or evaluated as an autonomous site-selection system, regulatory-compliance tool, permitting instrument, authoritative public map, or replacement for professional GIS, planning, and legal judgement. All generated output therefore requires substantive review by qualified professionals before they can inform subsequent planning activities.

The demonstrator evaluated in this study was not developed *ad hoc* for the focus groups. It emerged from an earlier iterative prototyping trajectory reported in prior feasibility work (Peters et al., 2025). In that earlier phase, development proceeded through multiple design iterations in collaboration with policy staff and GIS professionals from the Province of Utrecht, using existing opportunity map, particularly for wind energy, as reference artefacts to guide successive refinements. This provenance is relevant because it clarifies that the demonstrator is the result of domain-grounded iteration rather than a standalone "demo," and it explains why expert participants could meaningfully evaluate it against an professional reference map. Across iterations, the system evolved from text-oriented retrieval and generation toward producing structured, machine-readable geospatial outputs that can be integrated more directly into GIS workflows. Earlier iterations emphasized generating code-like or markup-style spatial representations (including GML-style outputs), whereas the version evaluated here returns a structured geospatial output designed for visualization and inspection in QGIS. This design trajectory matters methodologically as it positions the demonstrator as a workflow

10.3389/fpos.2026.1904617

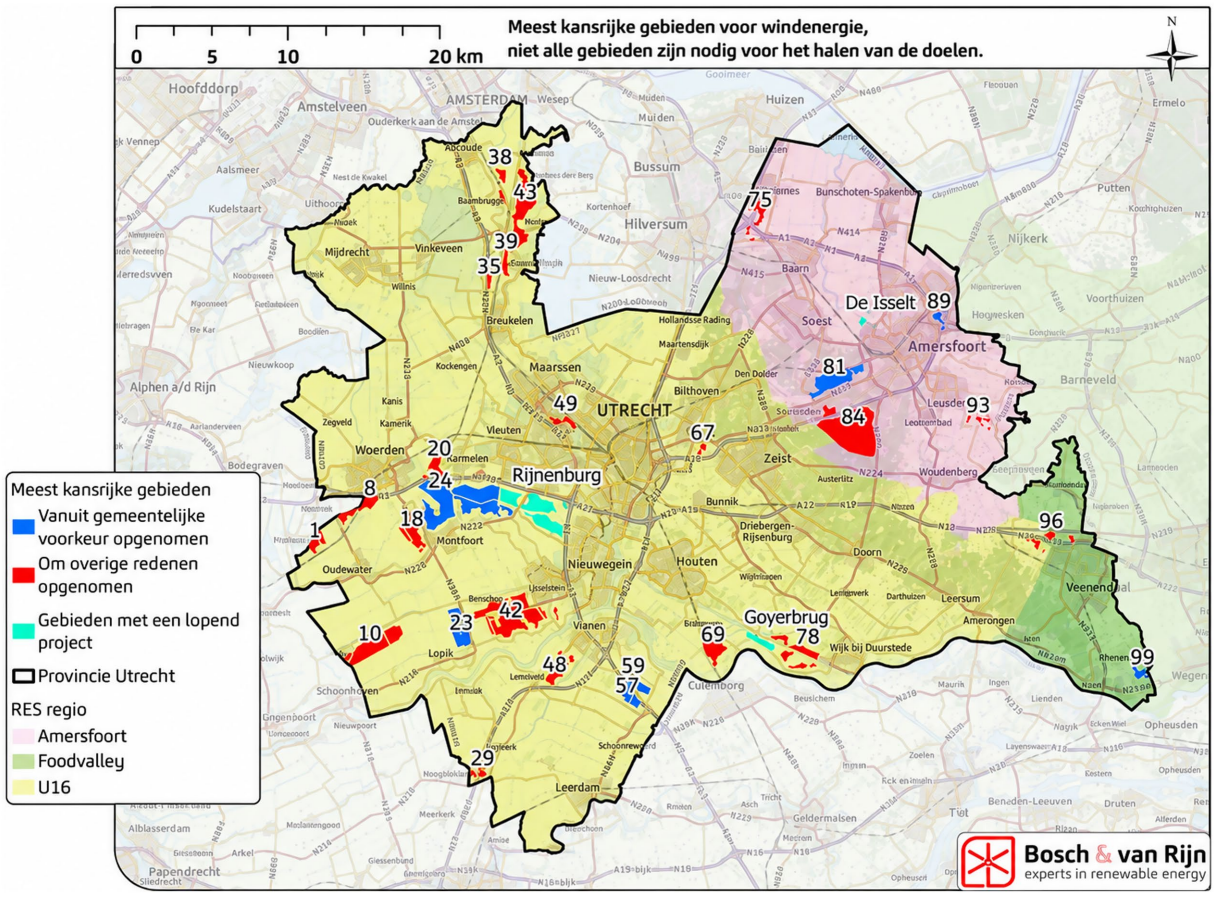

FIGURE 2
Professional reference opportunity map for wind-energy development produced by domain experts from the Province of Utrecht.

component for producing inspectable spatial planning artefacts, rather than as a generic conversational interface. Through iterative prompt tuning, additional rule-based guardrails were introduced to constrain model behavior and improve reliability, for example by enforcing that outputs cannot rely solely on geometric matches, must always include source references for legal claims, and must explicitly distinguish between formal rules, policy guidelines, and interpretations. These controls also included context-sensitive skill activation (e.g., only applying location-based reasoning when a query contains spatial references such as "this parcel" or coordinates), ensuring that responses remain grounded in both spatial context and legal traceability.

3.2.1 Demonstrator architecture and processing workflow

The demonstrator evaluated in this study was implemented as an orchestrated N8N workflow that combined generative and conventional processing components. The architecture was further developed from the earlier iterative feasibility study reported by Peters et al. (2025). Within the evaluated workflow, the generative component was used to interpret and structure information derived from policy and legal sources and to translate this information into a machine-readable representation for subsequent spatial processing and inspection, see Figure 3 (output) and Figure 4 (N8N architecture of the workflow). Rule-based guardrails required source references for legal claims and explicit differentiation between formal rules, policy guidance, and model interpretation. Context-dependent functions were activated only when relevant spatial information was present. The resulting structured geospatial artefact was subsequently visualized and inspected in QGIS.

The evaluated demonstrator used Google Gemini 2.5 Pro at temperature 0.7 with a 16,000-token output limit and otherwise default inference parameters. The model received role-based Dutch system prompts casting it as a spatial-planning and legal domain expert (reasoning agent) and as a quality controller (validation agent), each templated with the thematic object, its technical description, the retrieved legal context, a database schema reference, and the base polygon. Retrieval-augmented generation was used: policy and legal PDFs were chunked (~300 tokens, 50-token overlap), embedded with OpenAI text-embedding-3-large (3,072 dimensions), and indexed in FAISS (IndexFlatL2); the thirty highest-ranked chunks after location filtering were passed to the reasoning agent. Geospatial source data were accessed as a local GeoPackage (PDOK TOP10NL 2024, EPSG:28992) via GeoPandas/GDAL; the model itself received only a plain-text schema reference listing tables, columns, and their unique values, never the geometries. The LLM was responsible for interpreting the legal text, expressing each identified constraint as a table–column–value triple, classifying it internally as an absolute, conditional, or minor constraint, and justifying it. These internal labels correspond, respectively, to the domain-facing categories hard, complex, and soft used throughout the case evaluation. Deterministic processing performed attribute selection, geometry dissolution, iterative boolean difference from the base polygon, per-class aggregation, and export.

The intermediate output was a JSON array of {category, table, column, value, reasoning} objects, materialized as a multi-layer GeoPackage (remaining suitable area plus one excluded-area

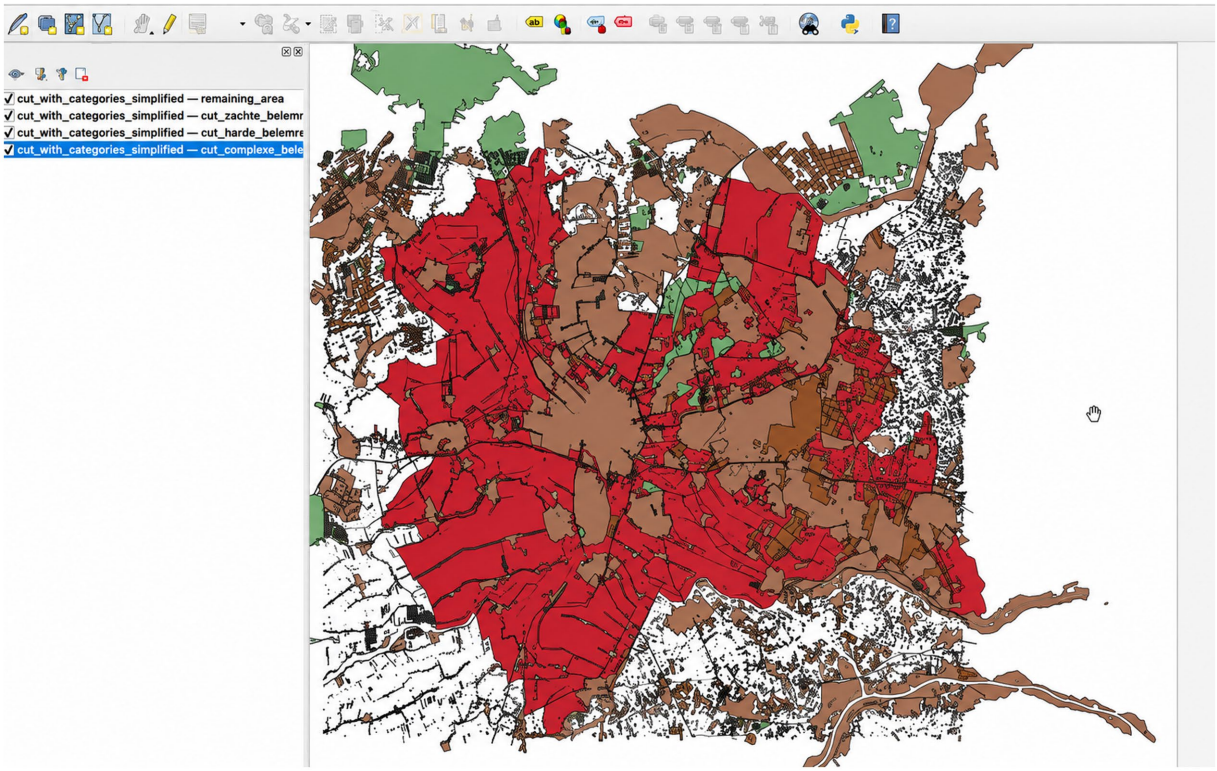

FIGURE 3
Demonstrator output presented in both FGs. The output presents the identification, mapping, and visualization of hard, complex, and soft constraints for wind-energy development in the Province of Utrecht.

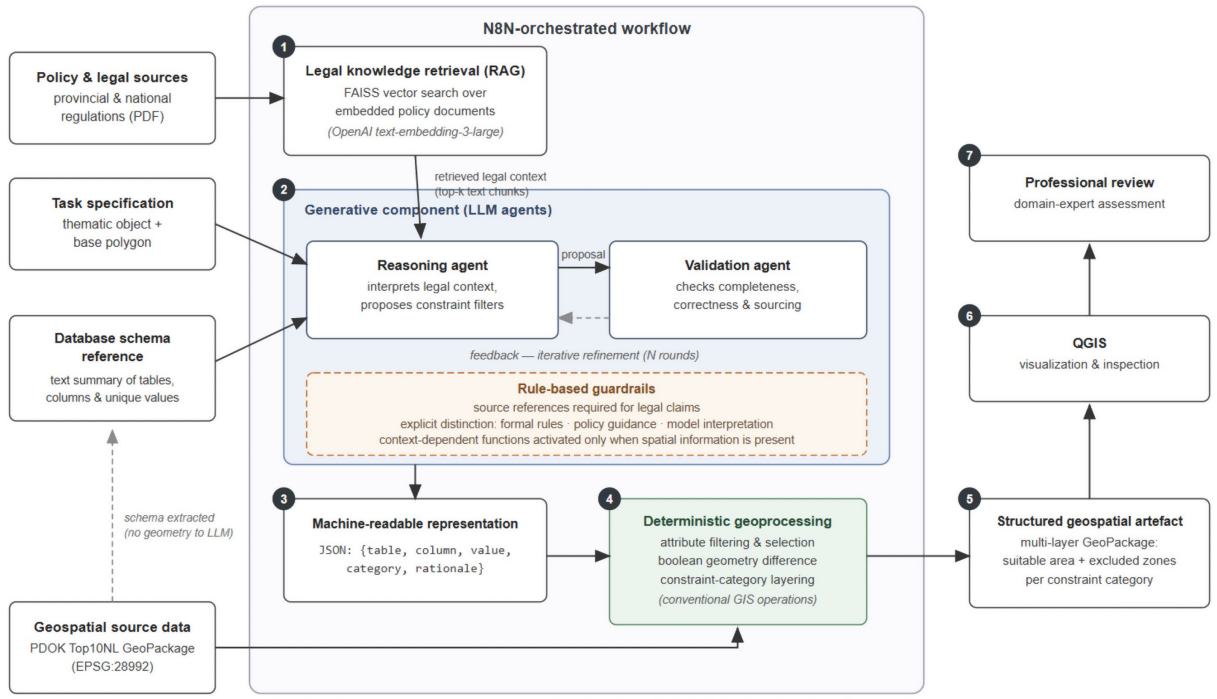

FIGURE 4
N8N orchestrated workflow of the demonstrator.

layer per severity class) before being loaded into QGIS. Automated validation comprised the validation agent's structured verdict driving up to two refinement rounds with early termination on approval, schema-aware parsing and type-checking of model output, and skipping of filters that resolved to empty selections; citation existence was not checked automatically. Professional review occurred after workflow completion, during inspection of the resulting GeoPackage in QGIS.

To further support interpretation and communication of the generated opportunity space, the demonstrator was extended with a 3D visualization step in which selected candidate locations from the AI-generated map were translated into the Netherlands3D environment. Based on official wind turbine specification data (e.g., hub height, rotor diameter, and total height) candidate turbines could be placed at real scale in relation to existing buildings, infrastructure, and landscape features. This made it possible to move beyond abstract 2D constraint mapping and instead explore the visual and spatial fit of potential turbine placements in a realistic environment, thereby supporting stakeholder dialogue and the presentation of alternatives in the planning process. An example of the resulting spatial 3D visualization is presented in Figure 5.

3.3 Expert FGs and participant selection

The data was collected using two purposively designed FGs. The use of complementary FGs, rather than a single unified group, reflects the recommendation that evaluations of complex sociotechnical PPSs benefit from deliberate separation of functional and technical knowledge domains, which may otherwise produce epistemic dominance by one stakeholder group and suppress (minority) perspectives (Brömmelstroet and Schrijnen, 2010; Geertman and Stillwell, 2020).

- FG1: Functional focus. FG was conducted in October 2025 and comprised 20 participants. Participants held roles spanning geo-information standardization, provincial GIS practice, policy implementation of the Dutch Omgevingswet, digital-twin strategy, and AI development and applied research. Organizations represented included Geonovum, multiple Dutch provinces (including Utrecht), the Ministry of Housing and Spatial Planning, Tygron, regional safety authorities, and the HU University of Applied Sciences Utrecht. Participant selection was purposive (Patton, 2015) as we recruited experts who either (a) create or formally validate spatial policy products such as opportunity maps, (b) manage or standardize geo-information infrastructures relevant to the Dutch Omgevingswet and the Digital System for the Physical Environment (DSO), or (c) develop or operationalize digital-twin and AI components within governmental contexts related to spatial planning.
- FG2: Technical focus. FG2 was conducted in November 2025 and included GIS specialists and AI practitioners from provincial authorities, Geonovum, HU University of Applied Sciences Utrecht, and regional safety bodies. The session did not constitute a formal quantitative spatial-accuracy assessment or software certification as this FG was specifically designed to provide an expert appraisal of the demonstrator's apparent legal, geometric, and technical alignment with professional practice and to identify requirements for more systematic validation, monitoring, and potential future operationalization.

3.4 Procedure and data collection

Both FGs followed an identical four-stage procedure to ensure comparability: (1) introduction and scoping, establishing shared understanding of the prototype and the wind-energy case context; (2) live demonstration of the full AI-assisted workflow, including the QGIS visualization and side-by-side comparison with the reference map produced by the provincial GIS professional; (3) guided expert discussion using a semi-structured protocol; and (4) administration of the TAM questionnaire immediately after the demonstration, before plenary discussion of results. FG1's discussion was structured around: (1) alignment of the demonstrator's output with the formal planning-policy cycle, (2) interpretability of AI-generated constraints and classifications for non-technical policy analysts, and (3) governance risks associated with the tool implicitly encoding or privileging particular policy interpretations without explicit deliberation. FG2's discussion was structured around: (1) geometric correctness of buffer zones and constraint classifications relative to authoritative datasets, (2) requirements for reproducibility of outputs across model runs under identical inputs, and (3) the feasibility of embedding structured validation steps, such as automated cross-checks against cadastral and regulatory datasets or integration of

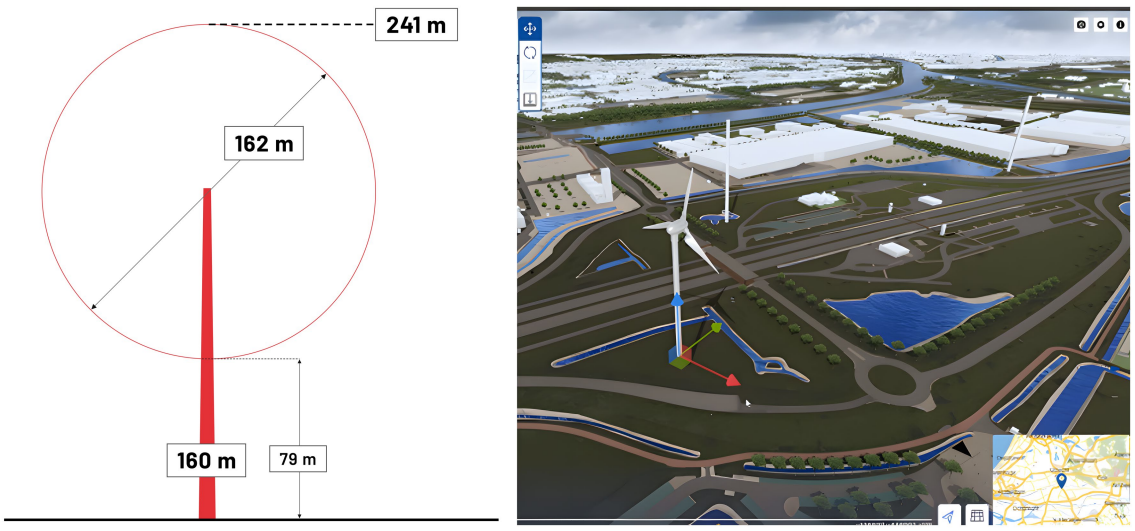

FIGURE 5
Translation of a candidate location from the AI-generated opportunity map into the Netherlands3D environment (Netherlands3D, 2026), using official wind turbine specifications (Verweij et al., 2023) to visualize the turbine at real scale in relation to surrounding buildings and landscape.

complementary AI agents including knowledge-graph and aerial-photography verification components. Reproducibility was addressed as an expert discussion topic and requirement for future technical validation. Repeated model runs under controlled identical inputs were not performed as part of this study.

In addition to the focus-group discussions, the generated opportunity map was subjected to a qualitative functional validation by a provincial domain expert. The expert compared the AI-generated wind-energy opportunity map with the existing professional reference case produced within the Province of Utrecht. The comparison focused on the treatment of constraint categories, the presence or absence of relevant buffers, the appropriateness of included spatial objects, the grounding of constraints in policy and legal sources, and the resulting residual opportunity space. This validation was qualitative: no area-based overlap measures, Intersection-over-Union (IoU), or other quantitative spatial-accuracy metrics were calculated.

Data sources comprised audio recordings and detailed FG notes (subject to participant consent and institutional permissions), artefacts generated during FGs (including annotated outputs, identified discrepancies, and structured issue lists compiled by participants), a written functional validation of the generated opportunity map by a provincial domain expert, and TAM questionnaire responses using Microsoft Forms within the secured Enterprise environment of the University. The mixed character of the data collection is consistent with best practice for the evaluation of PSSs, in which isolated survey instruments are recognized as insufficient to capture the contextual and relational dimensions of expert judgement (Brömmelstroet and Schrijnen, 2010). The protocols utilized and data collected are available upon request.

3.5 Measures: TAM and value elicitation

Technology acceptance was measured using a questionnaire grounded in the Technology Acceptance Model (TAM; Davis, 1989), the most extensively validated theoretical framework for explaining and predicting user adoption of information systems. TAM posits that two primary constructs, Perceived Usefulness (PU) and Perceived Ease of Use (PEoU), jointly determine behavioral intention to use a system, with PEoU additionally influencing PU. Extensions of TAM to AI-enabled and expert-facing systems confirm that transparency and explainability function as critical mediators between system features and trust-based acceptance in high-stakes institutional settings (Venkatesh and Bala, 2008; Wanner et al., 2022). Survey items were adapted from validated TAM operationalization's (Davis, 1989; Venkatesh and Bala, 2008) to reflect the spatial planning decision-support context, for example, referencing the accuracy of generated opportunity maps, the interpretability of constraint classifications for policy analysts, and the perceived utility of the workflow for realistic planning tasks. Responses were captured on a seven-point Likert scale (1 = strongly disagree; 7 = strongly agree). The same core TAM items were administered in both FGs to provide a common descriptive point of reference. Results from FG1 and FG2 are reported separately. Because the groups were purposively composed to represent different forms of expertise and are neither matched nor statistically representative samples, differences between their response patterns are interpreted cautiously and are not treated as evidence of population-level differences in technology acceptance.

Alongside the descriptive TAM measures, VSD was used primarily as an empirical and interpretive framework for eliciting public and professional values and translating stakeholder concerns into requirements for subsequent development (Friedman et al., 2017; Friedman and Hendry, 2019). Value elicitation was conducted through expert focus groups, which are well-suited for surfacing and refining stakeholders' underlying evaluative criteria (including value-laden requirements) in complex socio-technical settings (Keeney et al., 1990; Manzano, 2022). The evaluated demonstrator had already undergone an earlier technical development trajectory before the structured value elicitation reported in this study. The present evaluation therefore does not claim that all values identified in the focus groups were already technically operationalized in the evaluated artefact. FG1 elicited values explicitly through a structured value-ranking exercise and governance-oriented prompts, whereas FG2 surfaced related normative concerns more indirectly through discussions of validation, transparency, controllability, reproducibility requirements, architecture, and legal defensibility. In VSD terminology, the focus groups are therefore treated primarily as empirical investigations, because they examine how stakeholders perceive the technology and articulate values and value-related requirements, rather than as technical investigations of how specific technological properties operationalize those values.

These value articulations were then captured in analysis through a dedicated "normative/value" code register (e.g., transparency and explainability, controllability and human responsibility, fairness/collective interest) so that TAM acceptance data could be interpreted as value-conditioned rather than purely instrumental In FG1, participants were presented with a predefined list of 32 public values derived from the CODIO model (Meijer and Ruijer, 2021). Participants were asked to select the five values they considered most important in relation to the demonstrator and rank these five values in order of importance. They were not required to rank all 32 values. Participants could additionally propose a value if they considered an important value to be missing from the predefined list. Fourteen valid ranking responses were available for analysis. "Total mentions" refers to the number of respondents who included a value anywhere in their top five, whereas "first-priority mentions" refers to the number of respondents who ranked that value first.

3.6 Qualitative analysis and coding tree construction

Expert discussion data was analyzed using reflexive thematic analysis (Braun and Clarke, 2006, 2019), which is a theoretically flexible approach that supports both inductive emergence and deductive structuring of themes. This method was selected because the central evaluative concepts, readiness, trust, governance fitness, are simultaneously empirically grounded and normatively loaded, requiring a mode of analysis that can move between descriptive and interpretive approaches. The analytical process followed six phases: (1) data familiarization through repeated engagement with recordings and FG notes; (2) initial coding of concrete, observable noteworthy codes; (3) theme development by grouping codes into higher-order patterns; (4) theme review against the full dataset; (5) theme definition and naming; and (6) production of a structured coding tree. The coding tree integrated three code registers. First, descriptive codes capturing specific functional and technical deficiencies identified during the demonstration (e.g., missing buffer specifications, ambiguous classification of legal constraint categories, misalignment with the formal policy-cycle stage at which the

output would be used). Second, analytical codes capturing requirements for operationalization (e.g., provenance tracking, auditability of intermediate computational steps, human accountability for AI-informed spatial planning decisions, validation "checks and balances"). Third, normative codes capturing public-value concerns aligned with Dutch digital-government values frameworks and European AI governance guidance, specifically explainability, controllability, fairness of constraint weighting, and democratic legitimacy of AI-generated planning products (European Commission, 2021). Where multiple researchers participated in coding, a trail of codebook iterations and resolutions of coding disagreements was maintained in accordance with established criteria for qualitative rigor (Braun and Clarke, 2019; Lincoln and Guba, 1985). An example of results of the coding process could be found in Table 2.

3.6.1 Quantitative analysis and integration

The quantitative TAM data was analyzed descriptively and separately for each FG. Because of the small purposive samples and incomplete questionnaire response, no inferential statistical analyses were performed and no statistical comparison between FG1 and FG2 is intended. Percentages are therefore reported as descriptive summaries of the valid responses to individual items and, where relevant, are accompanied by absolute counts. The qualitative focus-group data was analyzed alongside these descriptive patterns to provide contextual insight into why participants perceived the demonstrator as potentially useful or problematic. The mixed-method integration is therefore interpretive rather than statistical (Creswell and Plano Clark, 2018). This reflects the established premise that readiness and adoption in complex professional domains are jointly shaped by demonstrated capability, perceived utility, and trust-related constraints that cannot be reduced to numerical scores alone (te Brömmelstroet, 2017; Veale and Brass, 2019).

TABLE 2 Example coding and observations.

Main theme	Subtheme	Observation
Organization	Adoption	"Stay very close to the user who will have to use it in their work, because it is disruptive, but it does not mean that the job will disappear; they need to become enthusiastic and feel ownership."
Technology	LLM limitations	"You are still dealing with a language model, so all the instabilities of language models are also present here."
Data foundation	Data governance	"Data governance is often not really established in many organizations that would like to integrate. No standardization, I can tell you that."
Substantive frameworks	Spatial boundaries	"You basically select the objects, but then there should actually be a buffer around them. It does not find some things, and we need to do something about that."

3.6.2 Ethics and governance considerations

This study involved adult professional participants who took part in expert focus groups and questionnaires. Participants were recruited and their data were handled in accordance with applicable institutional research-ethics norms (Koninklijke Nederlandse Akademie van Wetenschappen, 2018). Participation was voluntary and informed consent was obtained prior to data collection. Audio recording was subject to explicit per-FG permission. Given the public-sector setting and the potential decision-relevance of AI-generated spatial planning outputs, the expert discussions were explicitly structured to surface governance and public-value concerns as empirical findings in their own right. Participants of FG 1 were prompted using themes derived from Dutch digital-government values frameworks and European AI governance guidance (European Commission, 2021), addressing: (1) explainability of AI-generated outputs for non-specialist policy audiences, (2) carefulness and accuracy of spatial constraint classifications relative to authoritative legal definitions, (3) auditability and controllability of the workflow and its intermediate computational outputs, and (4) human accountability for final spatial planning decisions in which AI-generated artefacts play an evidentiary or analytical role. These dimensions are reported and interpreted alongside the TAM results in Section 4.

4 Results

The empirical results reported in this section are based on two complementary data sources. First, the structured questionnaire data comprise 14 completed responses from FG1 and 9 completed responses from FG2. The instrument used in FG1 captures technology-acceptance judgments and ranked public values, whereas FG2's instrument captures a technically oriented assessment through a combination of closed Likert-type items and open technical questions. Across both questionnaires, respondents represent a mix of public-sector, private-sector, and applied research backgrounds, with substantial professional experience in planning, geo-information, AI, and related domains. The respondents in FG1 reported a median of 15 years of work experience, while the respondents in FG2 reported a median of 5 years.

4.1 Qualitative expert comparison with the professional reference map

The first layer of evaluation consisted of a qualitative expert comparison between the demonstrator output and the professionally produced provincial reference map. As part of this comparison, a provincial domain expert conducted a functional validation of the generated wind-energy opportunity map against the existing professional reference case. The assessment focused on substantive discrepancies in constraint selection and classification, the application of buffers, the inclusion of spatial objects, the grounding of constraints in policy and legal sources, and the resulting residual opportunity space. The evaluation was deliberately qualitative and did not include quantitative geometric accuracy measures such as total area differences, areal overlap, Intersection-over-Union (IoU), or class-level accuracy. The findings in this section should therefore be interpreted as a qualitative expert assessment of substantive planning

discrepancies rather than as a quantified measure of spatial agreement. The most important discrepancies concerned the treatment of constraints. In the category of hard constraints, the prototype did not adequately distinguish between building types and appears to have included all buildings rather than separating residential or otherwise sensitive or incompatible functions (e.g., churches, petrol stations, schools, hospitals, daycare centers and care homes). In addition, buffers around housing were missing, even though such buffers form an implicit and obvious part of the normal professional workflow. The resulting map also appeared to include more green areas than only those legally relevant to exclusion, such as Natura 2000 or goose resting areas. For complex constraints, similar problems were observed. Important buffers were absent, and some infrastructure features appeared in categories where the GIS professional would not normally place them. The category of "soft constraints" was also found to be weakly grounded in the underlying policy documentation, suggesting that the demonstrator still translated parts of the policy material into spatial classes in ways that do not yet align with established domain interpretation. These discrepancies can be interpreted as domain-specific generative AI hallucinations, in which the model produces outputs that are internally coherent but misaligned with legal definitions, spatial logic, or implicit expert knowledge used in professional planning practice.

The identified discrepancies affect the substantive validity of the generated opportunity space. Missing buffers, insufficient differentiation of building and constraint types, and weak grounding of some policy-based constraints resulted in an opportunity area that was judged too broad from a professional planning perspective. The output was therefore structurally recognizable as an opportunity-map artefact, but it was not sufficiently aligned with the professional reference workflow to be treated as a reliable planning map. In particular, the identified errors would be unacceptable if the output were used as an authoritative representation of regulatory or planning constraints. The domain expert reached the same overall conclusion when evaluating the residual opportunity space: the generated remaining area was considered substantially too broad, which was attributed primarily to the selected constraint interpretations and the absence of relevant buffer zones. The comparison nevertheless provides useful diagnostic evidence about where further technical development and validation are required.

4.2 FG1 results: functional and governance expert evaluation

4.2.1 Technology acceptance (of the demonstrator in context)

Fourteen FG1 participants provided valid questionnaire responses in FG1. Descriptively, responses were more positive for items concerning the potential usefulness of the demonstrator than for its immediate ease of use or fit with current work practices. The detailed item-level response distributions are presented in Figure 6.

This contrast between perceived usefulness and limited perceived ease-of-use is substantively important. It suggests that the demonstrator is being valued primarily for what it could enable in planning practice rather than for its present form as a frictionless end-user tool. The structured data from FG1 therefore supports a nuanced interpretation of early acceptance. The participants see promise, but that promise is not yet accompanied by confidence that the current system would naturally fit into daily work processes without further development.

In addition to the structured data collection, the open-ended discussion in FG1 provided insights into how experts evaluated the demonstrator's usefulness, limitations, and practical fit within spatial planning work. Participants emphasized the demonstrator's potential usefulness, especially for early-stage scoping. One participant described it as "*a tool to estimate the likelihood that an environmental plan can be adjusted… and I find that very helpful.*" Others highlighted the speed advantage, noting that the demonstrator produced an answer "*in 30… eventually 45 seconds… whereas normally we need an hour, an hour and a half.*" Several participants added that the demonstrator achieve, albeit in limited for, what digital-governance ambitions have long promised: "*this is what the environmental-law digital system was supposed to do: ask 'what is allowed here?'*

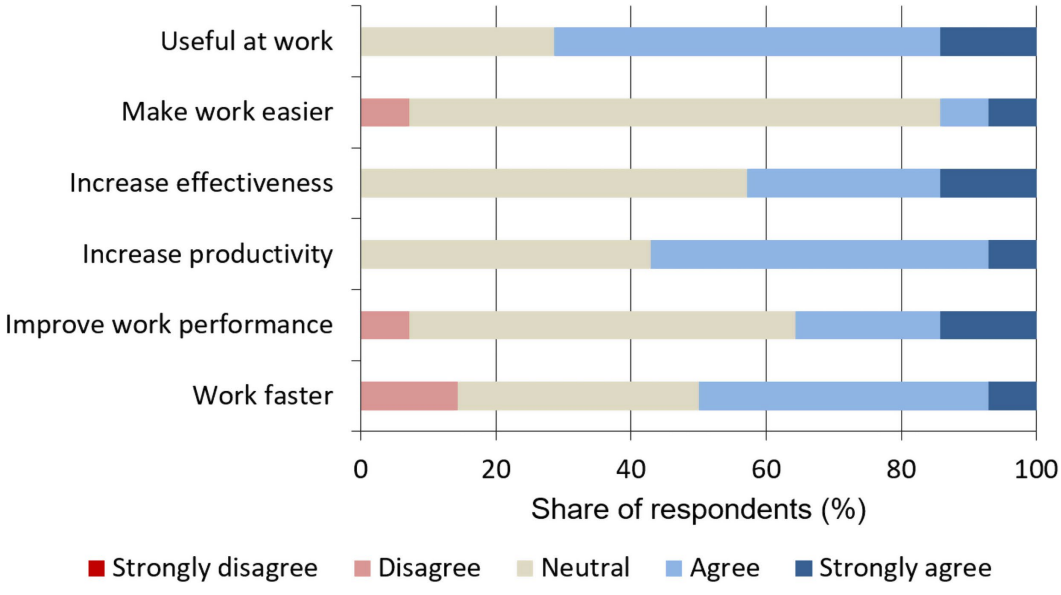

FIGURE 6
FG1 acceptance profile of the demonstrator. Shares of respondents on the TAM-items. Percentages are based on valid responses per item.

and get the answer. Here you see a working demo." Others reacted even more strongly: "*this is really awesome what we are doing… you can see it in the result, and that is very cool*." These comments strengthen the finding that perceived usefulness is the demonstrator's strongest acceptance dimension. At the same time, participants raised concerns regarding usability and practical integration. One participant stressed the need for staff training: "*you need to train them… train them in prompting… train them in thinking about the limitations of the prompt*." Another warned that the actual planning workflow is considerably more complex than the simplified process shown: "*the process is not one step… and you are trying, for simplicity, to make it one step*." These remarks reinforce the quantitative finding that ease-of-use was viewed far more cautiously.

A dominant theme concerned trustworthiness, correctness, and explainability. Participants worried that the model's generated reasoning could give a false sense of transparency. As one put it: "*it looks as if the chain-of-thought is the same as explainability, but it is not*." Another added that "*there may be a gap between what a model says it is doing and what is actually happening*." One participant expressed uncertainty about reliability, asking whether the system is correct "*90 percent of the time… or 98… or 70*." These concerns suggest that participants require validation procedures and a more inspectable decision logic before fully trusting the system. Participants also noted risks and unintended uses, especially if the tool were made widely accessible. One participant warned that "*land speculators would certainly be very interested… and is that what you want?*" Another highlighted the political risks of premature public visibility: "*citizens think in black-and-white terms, especially when they see maps… and then you create resistance yourself*." These remarks indicate that acceptance is tied not only to model performance but also to governance design and communication & citizen participation practices. Several participants emphasized user-group fit, noting that the demonstrator is primarily suited for policy professionals rather than broad public audiences. One said: "*you are making this for policy officers… so frame the policy process you want to support carefully*." Another warned that if the tool is positioned too broadly, "*you end up not being able to answer anything at all*." This indicates that current acceptance is closely linked to scoping, role clarity, and alignment with professional workflows.

Finally, participants reflected on the demonstrator's anticipated impact. Some were optimistic, calling it "*a game changer*." Others highlighted its potential to democratize access to spatial planning related information: "*the knowledge that currently lies with a small group of specialists could become available to many more people*." One participant emphasized generational differences, recounting that procedural GIS workflows were considered "*boring, and you're a boomer*." These comments indicate optimism about long-term adoption, provided that development continues and its concerns are addressed adequately and validated further.

4.2.2 Public values in relation to the demonstrator

The FG1 questionnaire also asked respondents to rank the public values they considered most important in relation to the demonstrator. Figure 7 reports the individual value labels as selected by respondents. No formal clustering of these labels was performed. Explainability, transparency, and controllability were among the more frequently selected values, while respondents also prioritized values concerning correctness, fairness, collective interests, and the prevention of harm. The distribution indicates that participants evaluated the demonstrator not only in terms of instrumental performance but also in relation to governance and public-value considerations.

These rankings show that respondents did not primarily understand the demonstrator as an instrumental efficiency tool. Instead, the dominant value pattern is primarily governance-oriented. Respondents prioritized whether the system is understandable, careful, controllable, and aligned with public interests over whether it is merely fast or technically impressive. This interpretation is reinforced by the open additions to the value list, where one respondent explicitly identified the risk of use for land speculation, and another highlighted correctness, defined as whether the reasoning and conclusions generated by the system are actually valid. Such additions show that

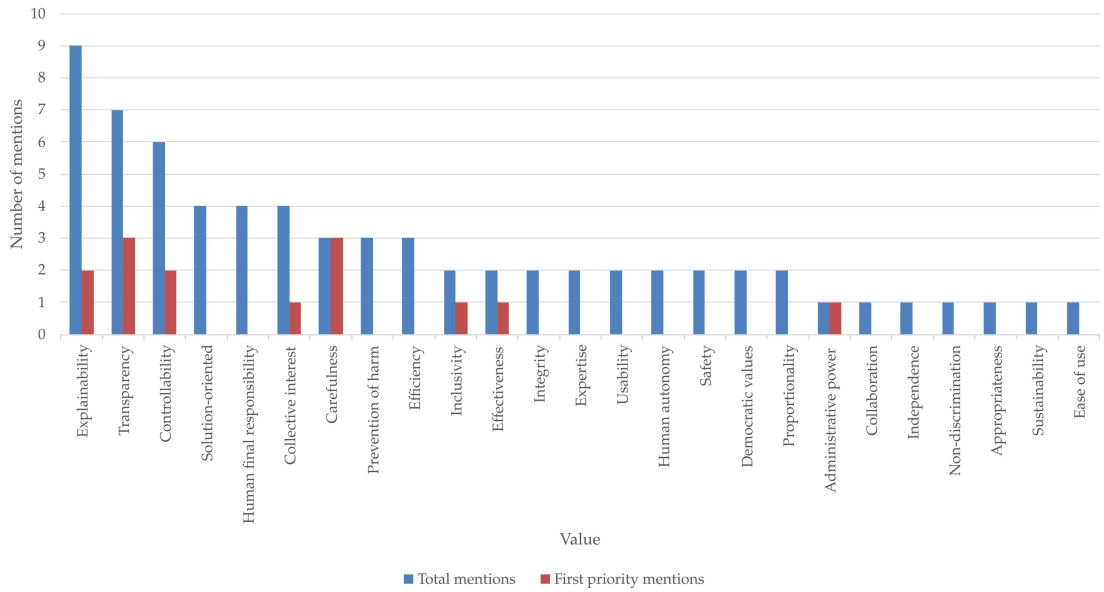

FIGURE 7
Public-value ranking in FG1. Total mentions indicate the number of respondents who included a value in their top five, whereas first-priority mentions indicate the number of respondents who ranked that value first. Participants selected and ranked five values from a predefined list of 32 CODIO-derived public values and could additionally propose a value they considered missing from the predefined list.

concerns about indirect harms and the reliability of model reasoning were already salient within the structured questionnaire itself.

Participants also reflected in the open-ended discussion on how core public values should be operationalized in the demonstrator's design and use. A recurring message was that the system must actively uphold foundational governance values, one participant warned that "*if it does not meet values such as transparency, openness, and so on, then we are missing the point*." Others emphasized that the development trajectory itself must reflect public-sector ethics, asking "*how do you ensure… that you serve the public interest?*" Participants also stressed that tools like this must reinforce, rather than undermine, institutional responsiveness and user trust, noting that meaningful use depends on "*staying very close to the actual users… because it is disruptive, but it is not as if their jobs will disappear*."

4.3 FG2 results: technical expert evaluation

4.3.1 Perceived usefulness and readiness in the technical session

Nine valid questionnaire responses were available for FG2s items. The descriptive response pattern indicates that participants saw potential in the demonstrator, particularly regarding speed and effectiveness, while responses concerning its immediate usefulness in professional practice were more mixed. Figure 8 presents the item-level distributions. These results are not used to statistically compare FG2 with FG1; rather, they provide a descriptive complement to the technical discussion reported further in this paper.

4.3.2 Open-ended technical discussion: readiness, validation, and implementation conditions

The open-ended discussion in FG2 adds a layer to the structured acceptance results by showing how technical experts interpreted the demonstrator's current maturity. A recurring theme was that the prototype was appreciated as a strong technical starting point, but not yet as a robust application. Participants repeatedly framed it as "*really still a proof of concept*" and, in one case, as "*a working garage version*." At the same time, the tone was positive. The prototype was seen as "*a very nice starting point*" and as something with "*potential*" that is worth developing further. In other words, the discussion did not question whether the demonstrator is promising, but rather whether it is sufficiently mature for broader operational use.

A second major theme concerned validation and failure analysis. Participants stressed that the central technical question is not only whether the system can generate plausible output, but also when, where, and why it fails. One participant explicitly stated that the key issue is being able to determine "*where it goes wrong*," and argued for much more systematic testing across use cases and input variations. Others suggested that future development should move parts of the reasoning away from the language model and into more explicit structures such as tables, decision trees, or tool calls, so that difficult policy interpretations and edge cases can be checked more reliably. This emphasis on validation is also visible in the open-ended questionnaire responses, which mention legal checks, qualitative evaluation across multiple use cases, and the need to identify disagreement or error sources before any scaling can be considered. Participants also linked technical readiness to architecture and system design. Several comments pointed to the need for stronger standards and a more modular setup, including references to Model Context Protocol (MCP), Docker, Kubernetes, model-agnostic design, and more robust geodata interfaces. One participant argued that the current configuration relies too heavily on free-form language-model reasoning and that more deterministic routes should gradually be introduced where possible. Others highlighted concrete improvement areas such as performance of geospatial operations, rendering, better data interfaces, more complete data input, and improved compatibility across different GIS environments. These remarks suggest that technical acceptance depends not only on the demonstrator's outputs, but also on whether the surrounding architecture can support reliability, scalability, and maintainability.

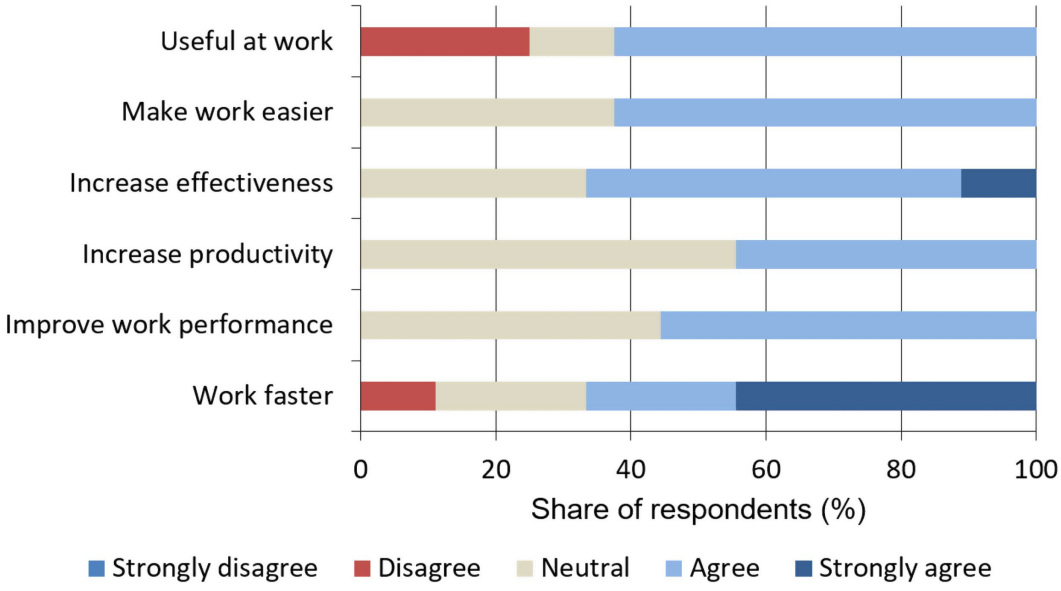

FIGURE 8
FG2 acceptance profile of the demonstrator. Shares of respondents on the TAM-items. Percentages are based on valid responses per item.

A final theme was the importance of competencies and user support. Technical experts consistently argued that meaningful use of the demonstrator requires a combination of AI literacy, Python and software-engineering knowledge, and geospatial/domain expertise. Several respondents stressed that users must understand both the possibilities and the limits of the underlying techniques, while others noted that some practical barriers for non-technical users could be mitigated through a better frontend, more transparent intermediate steps, and clearer documentation. Taken together, the open-ended discussion shows that FG2 respondents saw value in the demonstrator, but framed that value as conditional on validation, architectural refinement, stronger data integration, and the development of a broader support ecosystem around the prototype.

5 Discussion

Across the two FGs, participants perceived the demonstrator as an interesting and potentially useful direction for spatial-planning support, but the qualitative reference comparison and technical discussion also show that the evaluated output was not sufficiently dependable for routine or authoritative planning use at the current level of system maturity. The distinction between conceptual promise and output reliability is therefore central to interpreting the findings. The system produced a recognizable spatial artefact and demonstrated how policy and spatial information could be combined rapidly, but the observed errors in buffering, classification, and policy interpretation require substantive expert verification and correction.

The two FGs highlight different aspects of this early maturity state. FG1 primarily identified potential functional value and governance concerns, whereas FG2 emphasized the technical and architectural conditions that would need to be satisfied before such potential could translate into dependable use. This distinction is consistent with the long-standing planning-support literature, which shows that technical capability does not automatically translate into successful embedding in professional workflows and institutional practice (Geertman and Stillwell, 2004; Brömmelstroet and Schrijnen, 2010).

This maturity interpretation is also consistent with more recent work on digital twins and AI-enabled spatial planning, which frames such systems not as neutral mirrors of reality, but as selective socio-technical representations whose usefulness depends on interoperability, governance, transparency, and institutional embedding (Batty, 2024; Haraguchi et al., 2024; Therias and Rafiee, 2023). The present findings suggest that the demonstrator should be understood in these terms: not as a near-finished application that only requires engineering polish, but as an early planning-support component whose practical value depends on alignment between data, architecture, human judgement, and public-value safeguards. Viewed against the longer development of AI in urban planning described by Shabani et al. (2025), the present case can be understood as an empirical example of the broader shift from primarily computational planning tools toward more integrated socio-technical AI-enabled planning support. The findings particularly illustrate why this development cannot be assessed through technical capability alone: professional judgement, institutional embedding, validation, and governance remain integral to whether AI-supported planning artefacts can become meaningful in practice.

A second important finding is that perceived usefulness was stronger than perceived ease of use, particularly in FG1. In TAM terms, respondents recognized potential usefulness, productivity, and effectiveness, but were less convinced that the current demonstrator fits daily work without additional effort, training, and organizational redesign (Davis, 1989; Venkatesh and Bala, 2008). This indicates conditional acceptance as usefulness is acknowledged in principle, while dependable use in practice remains contingent on validation, clearer reasoning structures, and workflow integration. In public-sector AI contexts, this pattern should be interpreted more broadly than in conventional information-systems settings, because acceptance is mediated by trust-related variables such as transparency and explainability, which shape whether usefulness is judged as appropriate reliance in high-stakes settings (Grimmelikhuijsen, 2023; Wanner et al., 2022). This interpretation fits the present results closely. The demonstrator is not rejected, but its usefulness is treated as conditional on mechanisms that support verification, contestability, and accountability.

At the same time, qualitative remarks suggest that perceived ease of use may shift as interaction patterns evolve. Several participants emphasized the potential of dynamically adjusting inputs and immediately observing changes in outputs as a fundamentally different mode of working than traditional GIS workflows. This (near) real-time responsiveness introduces a form of interactive steering in which planners iteratively explore and refine constraints and relationships. This aspect is not fully captured by the current TAM responses, which reflect the demonstrator's present interface friction and maturity, but it suggests that future iterations should consider not only immediate usability but also emergent forms of interaction that could reduce cognitive and procedural burden when systems mature. The descriptive TAM results should be interpreted narrowly. The results indicate that several participating experts could envision useful applications for the concept, particularly in early-stage scoping, synthesis of complex information, and rapid exploration of spatial constraints. Broader professional acceptance was not established. Importantly, the qualitative findings show that perceived usefulness was repeatedly conditional on validation, clearer reasoning structures, workflow integration, and governance safeguards. Potential usefulness should therefore not be equated to readiness for professional reliance.

The value-related findings provide case-specific empirical support for an established pattern in public-sector and urban-governance AI research. Assessments of AI-supported decision tools are shaped not only by instrumental usefulness but also by concerns regarding transparency, accountability, fairness, controllability, legitimacy, and human oversight (Chen et al., 2023; Mišić et al., 2025; Zuiderwijk et al., 2021; Lartey and Law, 2025, 2026). In FG1, these concerns were articulated explicitly through the value-ranking exercise and governance-focused discussion. In FG2, related concerns emerged through technical discussions of validation, architectural robustness, standards, and legal defensibility. The present case therefore extends the empirical settings in which these socio-technical conditions have been observed, which is in line with the current body of knowledge.

The second FG makes clear that remaining barriers are not reducible to model capability alone. Respondents did not simply argue for "better LLMs," they emphasized validation workflows, modular architectures, standards compatibility, data interfaces, documentation, and user competence as prerequisites for adoption. This direction resonates with work on tool-augmented geospatial AI, where reliability improves when models are embedded in structured workflows with explicit task decomposition and deterministic checks rather than unconstrained

(large) language-based reasoning (Li et al., 2025; Van de Weghe et al., 2025). One architectural pathway for improving the inspectability of the workflow is to separate tasks such as policy interpretation, spatial transformation, and validation into explicit components. However, the findings also require a careful distinction between model-generated rationales and system-level explainability. As participants explicitly observed, a language model's account of its own reasoning should not automatically be treated as a faithful explanation of how an output was generated. Research on explanation faithfulness similarly distinguishes explanations that appear plausible to users from explanations that accurately represent the factors underlying a model's prediction (Jacovi and Goldberg, 2020), while empirical work on chain-of-thought prompting shows that generated reasoning can systematically misrepresent the factors that influenced a model's answer (Turpin et al., 2023). Explainability in this setting should therefore rely more strongly on externally verifiable artefacts, including provenance of legal and spatial sources, explicit representations of applied rules, inspectable tool calls, versioned intermediate outputs, and documented validation results.

Beyond efficiency gains, the findings also point to a broader form of cognitive augmentation. Participants noted that such systems may help reveal cross-domain relationships between spatial and policy domains (e.g., interactions between water systems, land use, and environmental constraints) that are difficult to identify through conventional manual analysis due to their complexity. At the same time, increased pattern detection and synthesis introduces governance challenges. Systems that generate large volumes of plausible interpretations may overwhelm deliberative processes, be strategically used to support particular interests, or induce automation bias, in which users place inappropriate reliance on imperfect automated decision support and reduce independent scrutiny (Parasuraman and Manzey, 2010). The errors observed in the present reference comparison illustrate why these risks are substantive. Incorrect exclusion or inclusion of locations could create false confidence about regulatory feasibility, while prematurely presenting generated opportunity maps to external stakeholders could influence political, commercial, or public interpretation before professional validation has occurred. These risks reinforce the need for explicit professional responsibility, documented validation procedures, transparency of assumptions, and mechanisms for contesting generated outputs. A final implication concerns the conditions under which such systems could move beyond prototype status, which was discussed to be a problem earlier in the paper. The findings suggest that future development will depend not only on improved model performance, but on the ability to embed generative AI within architectures that support validation, traceability, and institutional accountability. In particular, the requirement that planning outputs remain legally defensible in court implies that AI-supported workflows must be auditable and transparent at the level of intermediate steps, so that professional judgement can be exercised and decisions can withstand scrutiny in formal procedures. The key challenge is therefore not only technical advancement, but the development of socio-technical infrastructures that support responsible and legitimate use in planning practice.

Finally, it is worth noting one development reflection beyond the formal evaluation presented in this paper. Following the initial data collection, subsequent prototype iterations incorporated expert feedback (e.g., adjustments in constraint interpretation and output structuring). Informal observations during this process suggest that iterative refinement can reduce initial skepticism when users see their feedback reflected in system behavior, demonstrating acceptance as an evolving relationship rather than a one-time reaction. These post-evaluation observations were not measured systematically in the present study, but they align with the broader implication that trust and adoption, in this context, may emerge gradually through repeated cycles of validation, transparency improvements, and closer alignment with domain logic.

5.1 Limitations

Several limitations should be acknowledged when interpreting the findings. First, the empirical FGs involved relatively small numbers of respondents, which limits statistical generalizability. The study was deliberately designed around a highly specific expert group, being participants who create, validate, manage, or technically develop spatial-planning outputs. In that sense, the design prioritizes contextual depth over breadth. The findings should therefore not be read as representative of all spatial planning professionals or public-sector AI users, but they provide contextually rich expert evidence within this specific domain setting for informed further experimentation and development. A further limitation concerns the difference between the number of participants reported for the FGs in Table 1 and the lower number of questionnaire responses reported in the quantitative analysis. This discrepancy resulted from incomplete survey submissions and participants who chose not to participate filling in the questionnaire. These individuals nevertheless took part in the FGs and therefore contributed to the qualitative open-ended discussion. Consequently, the survey findings should be interpreted as representing the views of the valid respondents rather than the full number of FG attendees. This distinction is particularly important in a small-n expert design such as this study, because even modest differences between attendance and completed responses may affect the interpretation of response patterns. Consequently, the questionnaire findings should be interpreted solely as descriptive response patterns within the participating expert groups. The small numbers of valid responses do not support inferential claims about professional adoption, nor do they support statistical comparison between FG1 and FG2. The focus-group setting also introduces potential procedural limitations. Participants were purposively selected because of their expertise, the use of a live demonstration may have introduced novelty or demonstration effects, and the side-by-side presentation of the professional reference map may have anchored subsequent evaluation. Group interaction may additionally have influenced which concerns became most dominant during discussion. These factors are consistent with the exploratory character of the study but further limit the transferability of the observed response patterns (Manzano, 2022). Furthermore, the second FG did not employ a line-by-line code audit. What was assessed was the demonstrator as presented through its architecture, workflow logic, observable outputs, and selected implementation choices. This means that the technical findings should be read as an expert appraisal of technical credibility, risks, and implementation conditions, rather than as a software-engineering certification of robustness or maintainability. The technical discussion therefore provides contextual insight into technical maturity and readiness, but not a complete assessment of the code base.

A further limitation concerns the spatial reference comparison. Although the AI-generated opportunity map was compared with the professional reference case by a provincial domain expert, this validation remained qualitative. The study did not calculate quantitative geometric accuracy measures such as total opportunity-area

differences, areal overlap, Intersection-over-Union, or class-level spatial agreement. Consequently, the study provides evidence about substantive discrepancies identified through professional assessment, but does not provide a quantitative estimate of geometric similarity or spatial accuracy. Relatedly, reproducibility was identified by FG2 participants as an important technical requirement but was not experimentally assessed. The demonstrator was not run repeatedly under controlled identical inputs as part of this study. No conclusions can therefore be drawn regarding run-to-run stability or reproducibility of constraint classification and spatial output.

Another limitation concerns that the qualitative data were collected in Dutch and translated into English for reporting in the paper. Although the translations were carefully checked, some nuance may inevitably have been lost, especially in relation to governance language, public-value terminology, and the tone of open-ended comments. This issue is particularly relevant in a study where subtle distinctions between concepts such as transparency, explainability, carefulness, controllability, and correctness matter in an analytical sense. The reported findings should therefore be understood as carefully translated interpretations of the original data, while acknowledging the possibility of minor semantic loss.

We also think that the elicitation and interpretation of public values should be mentioned as a possible limitation of this study, which is often a complexity when working with VSD-like collection and analysis techniques (Friedman et al., 2006; Van Den Hoven et al., 2015). Working with values and their operationalization is always challenging when framing is insufficiently controlled, because respondents may understand the same value labels in different ways. In this study, that risk was mitigated by deliberately introducing and contextualizing the concept of public values during data collection and by grounding the work in a Value Sensitive Design (VSD) perspective. A further limitation is that public values were introduced and positioned explicitly only in the first FG. In the second FG, values were discussed more implicitly through themes such as validation, transparency, and controllability, which means that cross-FG comparison of values should be interpreted with caution. Finally, the role of VSD in the present study is primarily empirical and evaluative. The evaluated demonstrator was developed before the structured value elicitation reported here, and the study therefore does not demonstrate that all identified values were already technically operationalized in the evaluated artefact. Instead, the focus groups identify value-related requirements that can inform subsequent development cycles. Some existing design features can be examined in relation to values such as transparency and controllability, but the present study does not experimentally compare alternative value-sensitive designs (cf. Friedman et al., 2017; Friedman and Hendry, 2019).

6 Conclusion

This study examined an exploratory generative AI–based system for translating legal, policy, and geospatial information into a wind-energy opportunity map in the Province of Utrecht. The case demonstrates that a generative-AI workflow can rapidly produce a structured and recognizable geospatial artefact from heterogeneous planning information. At the same time, qualitative comparison with an existing professional reference map revealed crucial deficiencies, including missing buffers, insufficient differentiation of constraint types, and weak grounding of some policy-based constraints. These errors mean that the evaluated output cannot currently be relied upon as an authoritative planning or regulatory artefact without substantive expert verification and correction.

The expert evaluation nevertheless provides evidence that participants saw potential value in the underlying workflow, particularly for rapid exploration, synthesis, and early-stage planning support. This perceived potential was consistently conditional on stronger validation, traceability, workflow integration, controllability, and human oversight. The descriptive TAM data should be interpreted within the limited expert samples and do not establish broader professional adoption. Similarly, VSD functioned primarily as an empirical and interpretive lens for identifying normative requirements rather than demonstrating that these requirements were already technically operationalized in the evaluated system.

The contribution of the study is therefore deliberately localized. It provides an in-depth empirical case of how experts evaluate a generative-AI planning demonstrator, identifies concrete spatial and technical failure points, and shows how perceived usefulness is considered alongside governance and public-value concerns in this specific context. These observations are consistent with broader research showing that AI adoption in public and urban governance is shaped by institutional and socio-technical conditions alongside technical capability (Lartey and Law, 2025, 2026; Zuiderwijk et al., 2021).

The present evidence supports continued development and systematic validation rather than operational deployment. Before this type of system could support formal spatial-planning processes, future work needs to establish quantitative spatial accuracy, run-to-run reproducibility, reliable legal and policy-rule interpretation, auditable provenance, and clear mechanisms for professional review and accountability. Hybrid architectures combining generative components with deterministic geospatial and rule-based validation therefore represent an important next development step (Li et al., 2025; Van de Weghe et al., 2025; Zhang et al., 2025).

7 Future research agenda

The present study points to a future research agenda that is explicitly shaped by both the empirical findings and the development trajectory of the demonstrator itself. Much of the recent literature on generative AI in spatial planning has focused on conceptual analyses, systematic reviews, or isolated technical capabilities. The present study adds to that agenda by identifying, from the perspective of expert users and developers in context, where the main barriers to further development currently lie. These barriers are not purely technical. Future research should therefore treat generative AI in spatial planning as a socio-technical development challenge, not as a technical problem, such as model improvement, alone.

A first research avenue concerns technical maturation and validation by design. The findings show that future work should move toward architectures that incorporate stronger validation logic, structured tool use, and more explicit reasoning support. This includes research into hybrid workflows in which language models are combined with deterministic geospatial functions, rule-based checks, legal decision tables, and formalized retrieval mechanisms. Such work is particularly relevant for planning tasks in which every output must be

explainable, traceable, and legally defensible. The current demonstrator represents the highest feasible maturity level within the present development cycle, but the next step is not simply to scale it. It is to investigate how robustness, validation completeness, and reproducibility can be systematically built into future iterations.

Moving from an exploratory prototype toward formal planning support would require a staged validation pathway. A first stage should establish quantitative spatial accuracy against authoritative reference data. Such validation should quantify, at minimum, the total AI-generated and reference opportunity areas, their ratio, areal overlap, and Intersection-over-Union. Where constraint categories are semantically comparable between the demonstrator and reference workflow, class-level spatial agreement should also be assessed. A second stage should assess run-to-run reproducibility and robustness under controlled input variations. This should include repeated runs under identical inputs and reporting variation in extracted constraint categories, hard/complex/soft classification, generated spatial geometries, and residual opportunity area. A third stage should validate the interpretation and application of legal and policy rules and ensure provenance at each processing step. Subsequent controlled pilots should then examine professional sign-off, accountability, monitoring, and the conditions under which AI-generated artefacts could contribute to formal planning procedures. This progression is consistent with current work emphasizing structured tool use, deterministic geospatial operations, and explicit validation mechanisms in LLM-enabled geospatial systems (Li et al., 2025; Van de Weghe et al., 2025; Zhang et al., 2025). A second research avenue concerns professional competencies and human-AI collaboration.

The results indicate that meaningful use of this type of system requires more than generic AI literacy. Future research should examine what combinations of geospatial expertise, legal-policy understanding, data competence, and technical skills are needed to work responsibly with such tools in practice. In high-stakes spatial planning environments, the challenge is not merely whether people can operate the tool, but whether they can critically interpret, validate, and contest its outputs when necessary. A third research avenue concerns organizational fit and the wider sociotechnical environment. One of the clearest findings of this study is that respondents see promise in the demonstrator, but not yet a natural fit with existing work processes. Future research should therefore examine how AI-supported planning tools can be embedded in the highly complex, political, legal, and situation-specific environments in which spatial planning decisions are made. A fourth and particularly important research avenue concerns public values and VSD. The current study identifies value-related requirements that should inform subsequent demonstrator development and provides an initial basis for examining how existing technical design choices may support or constrain these values. However, because the current study does not compare multiple variants of the demonstrator in which values are operationalized differently, future research should move toward comparative VSD-oriented prototyping, in which alternative configurations of transparency, controllability, explanation, human oversight, or fairness are explicitly designed and then evaluated with different stakeholder groups. Such work would make it possible to study not only whether values matter, which the present study strongly suggests they do, but also how different value operationalizations are experienced in practice and what trade-offs they create across stakeholder perspectives. This also calls for the development of concrete mechanisms to operationalize these values in practice, including structured prompting where appropriate, knowledge graphs, retrieval-augmented generation, verifiable provenance records, explicit rule traces, inspectable tool calls and transformations, and versioned archival of model inputs, intermediate artefacts, and outputs.

A key open question emerging from this study concerns how such systems can be generalized into robust and reusable architectural frameworks. Future research should therefore investigate how agent-based, multimodal planning-support workflows can be formalized into generic architectures that support validation, traceability, and reproducibility across use cases. In particular, the certification of AI-supported planning outputs, both technically and legally, will be essential if such systems are to be used in formal decision-making contexts. This includes the ability to audit intermediate reasoning steps, verify compliance with legal frameworks, and ensure that outputs can withstand scrutiny in juridical settings. Addressing these challenges will be critical for moving from experimental prototypes toward institutionally embedded planning-support systems.

A final and particularly important research direction concerns the development of neuro-symbolic AI approaches for spatial planning support. The findings of this study, as well as broader literature on generative AI, indicate that limitations such as hallucinations and inconsistent reasoning are not specific to the spatial planning domain, but stem from the probabilistic and non-deterministic nature of neural models themselves. While approaches such as retrieval-augmented generation (RAG) can reduce these effects by grounding outputs in external data, they do not eliminate the risk of plausible but incorrect interpretations, especially in legally and spatially complex contexts. This points to future research that should move beyond standalone or retrieval-augmented LLM architectures toward hybrid neuro-symbolic systems, in which neural components are combined with explicit symbolic representations such as rules, knowledge graphs, and formal geospatial operations. In such architectures, neural AI can support the interpretation and synthesis of policy and legal information, while symbolic components provide deterministic reasoning, validation, and traceability. For spatial planning, where outputs must be traceable, explainable, reproducible, and legally defensible, this integration is likely to be a key condition for moving from exploratory prototypes toward operational decision-support systems.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements because Informed consent was given, only not written informed consent, by agreeing to join the data collection

sessions the participants gave informed consent to be involved in the study. No personal data was collected. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

KS: Conceptualization, Data curation, Formal analysis, Methodology, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. RP: Conceptualization, Funding acquisition, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. SL: Data curation, Formal analysis, Investigation, Methodology, Project administration, Validation, Visualization, Writing – review & editing. JM: Conceptualization, Data curation, Investigation, Resources, Visualization, Writing – review & editing. FK: Investigation, Software, Visualization, Writing – review & editing. RG: Software, Validation, Visualization, Writing – review & editing.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Acknowledgments

This research has been made possible by the Province of Utrecht GIS team. Especially Jasper Swager, Carolien Idema, Luc de Horde, Serhat Çaliscan, Erika Romijn, Marjolijn Gerdes, Bregtje Rijerman, Tim Stafleu, Arno Töpfer and Kirsten Blok. We also thank Bart de Lathouwer from Geonovum.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that Generative AI was used in the creation of this manuscript. We used generative AI solely to improve clarity and readability of the paper text. All ideas presented in the paper are original and originate from the authors.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Afroogh, S., Akbari, A., Malone, E., Kargar, M., and Alambeigi, H. (2024). Trust in AI: progress, challenges, and future directions. *Hum. Soc. Sci. Commun.* 11:1568. doi: 10.1057/s41599-024-04044-8

Batty, M. (1995) 'Planning Support Systems and the New Logic of Computation'. Available online at: https://www.semanticscholar.org/paper/Planning-support-systems-and-the-new-logic-of-Batty/735472225d6f5c6119bab94ad7c814cf086430aa (Accessed June 8, 2026)

Batty, M. (2024). Digital twins in City planning. *Nat. Comput. Sci.* 4, 192–199. doi: 10.1038/s43588-024-00606-7

Brama, H., Dalach, A., Grinshpoun, T., and Dortheimer, J. (2024) Towards a robust evaluation framework for generative urban design. In *Proceedings of the 42nd Conference on Education and Research in Computer Aided Architectural Design in Europe (eCAADe 2024)*

Braun, V., and Clarke, V. (2006). Using thematic analysis in psychology. *Qual. Res. Psychol.* 3, 77–101. doi: 10.1191/1478088706qp063oa

Braun, V., and Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qual. Res. Sport Exerc. Health* 11, 589–597. doi: 10.1080/2159676X.2019.1628806

Brömmelstroet, M. T., and Schrijnen, P. M. (2010). From planning support systems to mediated planning support: a structured dialogue to overcome the implementation gap. *Environ. Plan. B Plan. Des.* 37, 3–20. doi: 10.1068/b35019

Chen, Y.-C., Ahn, M. J., and Wang, Y.-F. (2023). Artificial intelligence and public values: value impacts and governance in the public sector. *Sustainability* 15:4796. doi: 10.3390/su15064796

Creswell, J. W., and Plano Clark, V. L. (2018). *Designing and Conducting Mixed Methods Research*. 3rd. SAGE Publications. Available online at: https://collegepublishing.sagepub.com/products/designing-and-conducting-mixed-methods-research-3-241842 (Accessed June 8, 2026).

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* 13, 319–340. doi: 10.2307/249008

De Lathouwer, B., Van Putten, G., Horbach, S., and Grothe, M. (2024) Policy Processes and Building Blocks for Digital Twins. Report Version 1.3. Geonovum Available online at: https://www.geonovum.nl/uploads/documents/Eindrapport%20Advies%20Beleid%20en%20Digital%20Twins%20-%20provincie%20Utrecht%20v1.3d_EN.pdf (Accessed June 8, 2026)

European Commission (2021) Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. COM(2021) 206 final. European Commission Available online at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 (Accessed June 8, 2026)

Friedman, B., and Hendry, D. G. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination*. Cambridge, Massachusetts: The MIT Press.

Friedman, B., Hendry, D. G., and Borning, A. (2017). A survey of value sensitive design methods. *Found. Trends® Hum. Comput. Interact.* 11, 63–125. doi: 10.1561/1100000015

Friedman, B., Kahn, P. H. Jr., and Borning, A. (2002) *Value Sensitive Design: Theory and Methods. UW CSE Technical Report No. 02-12-01.* University of Washington Available online at: https://dada.cs.washington.edu/research/tr/2002/12/UW-CSE-02-12-01.pdf (Accessed June 8, 2026)

Friedman, B. Jr., Kahn, P. H., and Borning, A. (2006). 'Value sensitive design and information systems'. In *Human-Computer Interaction in Management Information Systems: Foundations*, edited by P. Zhang and D. Galletta. Armonk, New York: M.E. Sharpe. Available online at: https://old.vsdesign.org/publications.php (Accessed June 8, 2026).

Geertman, S., and Stillwell, J. (2004). Planning support systems: an inventory of current practice. *Comput. Environ. Urban. Syst.* 28, 291–310. doi: 10.1016/S0198-9715(03)00024-3

Geertman, S., and Stillwell, J. (2020). Planning support science: developments and challenges. *Environ. Plann. B Urban Anal. City Sci.* 47, 1326–1342. doi: 10.1177/2399808320936277

Grimmelikhuijsen, S. (2023). Explaining why the computer says no: algorithmic transparency affects the perceived trustworthiness of automated decision-making. *Public Adm. Rev.* 83, 241–262. doi: 10.1111/puar.13483

Haraguchi, M., Funahashi, T., and Biljecki, F. (2024). Assessing governance implications of City digital twin technology: a maturity model approach. *Technol. Forecast. Soc. Chang.* 204:123409. doi: 10.1016/j.techfore.2024.123409

Jacovi, A., and Goldberg, Y. (2020) 'Towards faithfully interpretable NLP systems: how should we define and evaluate faithfulness?' In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* 4198–4205 Doi: 10.18653/v1/2020.acl-main.386

Jiang, H., Li, M., Witte, P., Geertman, S., and Pan, H. (2025). Urban chatter: exploring the potential of ChatGPT-like and generative AI in enhancing planning support. *Cities* 158:105701. doi: 10.1016/j.cities.2025.105701

Keeney, R. L., von Winterfeldt, D., and Eppel, T. (1990). Eliciting public values for complex policy decisions. *Manag. Sci.* 36, 1011–1030. doi: 10.1287/mnsc.36.9.1011

Klosterman, R. E. (1997). Planning support systems: a new perspective on computer-aided planning. *J. Plan. Educ. Res.* 17, 45–54. doi: 10.1177/0739456X9701700105

Koninklijke Nederlandse Akademie van Wetenschappen (2018). *Nederlandse Gedragscode Wetenschappelijke Integriteit. Gedragscode*. Amsterdam: Koninklijke Nederlandse Akademie van Wetenschappen.

Lartey, D., and Law, K. M. Y. (2025). Artificial intelligence adoption in urban planning governance: a systematic review of advancements in decision-making, and policy making. *Landsc. Urban Plan.* 258:105337. doi: 10.1016/j.landurbplan.2025.105337

Lartey, D., and Law, K. M. Y. (2026). Governing with artificial intelligence: mapping the knowledge systems shaping urban intelligence. *Technol. Soc.* 86:103321. doi: 10.1016/j.techsoc.2026.103321

Lavin, A., Gilligan-Lee, C. M., Visnjic, A., Ganju, S., Newman, D., Ganguly, S., et al. (2022). Technology readiness levels for machine learning systems. *Nat. Commun.* 13:6039. doi: 10.1038/s41467-022-33128-9

Li, Z., Ning, H., Gao, S., Janowicz, K., Li, W., Arundel, S. T., et al. (2025). GIScience in the era of artificial intelligence: a research agenda towards autonomous GIS. *Ann. GIS* 31, 501–536. doi: 10.1080/19475683.2025.2552161

Li, J., Qian, L., Liu, P., and Liu, T. (2024). Construction of legal knowledge graph based on knowledge-enhanced large language models. *Information* 15:666. doi: 10.3390/info15110666

Lin, Q., Hu, R., Li, H., Wu, S., Li, Y., Fang, K., et al. (2025). ShapefileGPT: a multi-agent large language model framework for automated shapefile processing. *Int. J. Digit. Earth* 18:2577884. doi: 10.1080/17538947.2025.2577884

Lincoln, Y. S., and Guba, E. G. (1985). *Naturalistic Inquiry*. SAGE Publications. Available online at: https://uk.sagepub.com/en-gb/eur/naturalistic-inquiry/book842 (Accessed June 8, 2026).

Mankins, J. C. (1995) Technology Readiness Levels: A White Paper. White Paper. NASA Office of Space Access and Technology Available online at: https://www.artemisinnovation.com/images/TRL_White_Paper_2004-Edited.pdf (Accessed June 8, 2026)

Manvi, R., Khanna, S., Mai, G., Burke, M., Lobell, D., and Ermon, S. (2024). GeoLLM: extracting geospatial knowledge from large language models. *arXiv* Available online at: http://arxiv.org/abs/2310.06213 (Accessed June 8, 2026).

Manzano, A. (2022). Conducting focus groups in realist evaluation. *Evaluation* 28, 406–425. doi: 10.1177/13563890221124637

Meijer, A., and Ruijer, E. (2021) *Code Goed Digitaal Openbaar Bestuur (CODIO): Borgen van Waarden Bij de Digitalisering van Het Openbaar Bestuur*. Utrecht: USBO Advies.

Mišić, J., Van Est, R., and Kool, L. (2025). Good governance of public sector AI: a combined value framework for good order and a good society. *AI Ethics* 5, 4875–4889. doi: 10.1007/s43681-025-00751-3

Netherlands3D (2026). 'Netherlands3D'. Available online at: https://netherlands3d.eu/ (Accessed June 8, 2026)

Neumann, O., Guirguis, K., and Steiner, R. (2024). Exploring artificial intelligence adoption in public organizations: a comparative case study. *Public Manag. Rev.* 26, 114–141. doi: 10.1080/14719037.2022.2048685

Pan, F., Huang, X., Bi, Y., Gao, Y., Ye, Y., and Wang, H. (2025). From tools to partners: how large language models are transforming urban planning. *AI Open* 6, 276–298. doi: 10.1016/j.aiopen.2025.11.001

Parasuraman, R., and Manzey, D. H. (2010). Complacency and Bias in human use of automation: an attentional integration. *Hum. Factors* 52, 381–410. doi: 10.1177/0018720810376055

Patton, M. Q. (2015). *Qualitative Research & Evaluation Methods: Integrating Theory and Practice*. 4th. SAGE Publications. Available online at: https://collegepublishing.sagepub.com/products/qualitative-research-evaluation-methods-4-232962. (Accessed June 8, 2026)

Pelzer, P., Geertman, S., Van Der Heijden, R., and Rouwette, E. (2014). The added value of planning support systems: a practitioner's perspective. *Comput. Environ. Urban. Syst.* 48, 16–27. doi: 10.1016/j.compenvurbsys.2014.05.002

Peters, R., Kok, F., Van Meerten, J., and Smit, K. (2025) Exploring the Feasibility of Generative AI in enhancing the identification of spatial and regulatory opportunities using urban digital twins. In *38th Bled eConference: Empowering Transformation: Shaping Digital Futures for All*. 745–754 doi: 10.18690/um.fov.4.2025.47

Rittel, H. W. J., and Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy. Sci.* 4, 155–169. doi: 10.1007/BF01405730

Sadek, M., Calvo, R. A., and Mougenot, C. (2024). Designing value-sensitive AI: a critical review and recommendations for socio-technical design processes. *AI Ethics* 4, 949–967. doi: 10.1007/s43681-023-00373-7

Shabani, A., AlWaer, H., and Taheri, S. (2025). AI revolution in urban planning: a 30-year journey–an overview and bibliometric analysis. *Proc. Inst. Civ. Eng. Urban Des. Plann.* 179, 41–57. doi: 10.1680/jurdp.25.00009

Simon, J., Wong, P.-H., and Rieder, G. (2020). Algorithmic bias and the value sensitive design approach. *Internet Policy Rev.* 9, 1–16. doi: 10.14763/2020.4.1534

te Brömmelstroet, M. (2017). PSS are more user-friendly, but are they also increasingly useful? *Transp. Res. Part A Policy Pract.* 104, 96–107. doi: 10.1016/j.tra.2016.08.009

Therias, A., and Rafiee, A. (2023). City digital twins for urban resilience. *Int. J. Digit. Earth* 16, 4164–4190. doi: 10.1080/17538947.2023.2264827

Turpin, M., Michael, J., Perez, E., and Bowman, S. (2023). Language models Don't always say what they think: unfaithful explanations in chain-of-thought prompting'. *Adv. Neural Inf. Proces. Syst.* 36, 74952–74965. doi: 10.52202/075280-3275

Van de Weghe, N., De Sloover, L., Cohn, A., Huang, H., Scheider, S., Sieber, R., et al. (2025). Opportunities and challenges of integrating geographic information science and large language models. *J. Spat. Inf. Sci.* 30, 93–116. doi: 10.5311/josis.2025.30.389

Van Den Hoven, J., Vermaas, P. E., and Van De Poel, I. (2015). *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*. Dordrecht: Springer Netherlands.

Veale, M., and Brass, I. (2019). "Administration by algorithm? Public management meets public sector machine learning," in *Algorithmic Regulation*, eds. K. Yeung and M. Lodge (Oxford: Oxford University Press).

Venkatesh, V., and Bala, H. (2008). Technology acceptance model 3 and a research agenda on interventions. *Decis. Sci.* 39, 273–315. doi: 10.1111/j.1540-5915.2008.00192.x

Verweij, W., Koopmans, B., and Maan, M. (2023).Notitie reikwijdte en detailniveau: PlanMER windenergie provincie Utrecht [Ontwerp]. Bosch & van Rijn. Available online at: https://pas.commissiemer.nl/files/nl/3719/011277-3719-nrd-planmer-windenergie-provincie-utrecht-ontwerp.pdf

Wang, Y., Fu, Y., and Lyles, W. (2026). Generative AI for spatial regeneration planning: integrating urban planning theories and ethics. *Comput. Environ. Urban. Syst.* 124:102380. doi: 10.1016/j.compenvurbsys.2025.102380

Wang, S., Hu, T., Xiao, H., Li, Y., Zhang, C., Ning, H., et al. (2024). GPT, large language models (LLMs) and generative artificial intelligence (GAI) models in geospatial science: a systematic review. *Int. J. Digit. Earth* 17:2353122. doi: 10.1080/17538947.2024.2353122

Wanner, J., Herm, L.-V., Heinrich, K., and Janiesch, C. (2022). The effect of transparency and trust on intelligent system acceptance: evidence from a user-based study. *Electron. Mark.* 32, 2079–2102. doi: 10.1007/s12525-022-00593-5

Wu, H., Shen, Z., Hou, S., Liang, J., Jiao, H., Qing, Y., et al. (2025). Autogeeval: a multimodal and automated evaluation framework for geospatial code generation on GEE with large language models. *ISPRS Int. J. Geo Inf.* 14:256. doi: 10.3390/ijgi14070256

Xu, H., Omitaomu, F., Sabri, S., Zlatanova, S., Li, X., and Song, Y. (2024). Leveraging generative AI for urban digital twins: a scoping review on the autonomous generation of urban data, scenarios, designs, and 3D city models for smart city advancement. *Urban Inf.* 3:29. doi: 10.1007/s44212-024-00060-w

Yang, A., Fu, C., Jia, Q., Dong, W., Ma, M., Chen, H., et al. (2025). Evaluating and enhancing spatial cognition abilities of large language models. *Int. J. Geogr. Inf. Sci.* 39, 2009–2044. doi: 10.1080/13658816.2025.2490701

Zhang, Q., Gao, S., Wei, C., Zhao, Y., Nie, Y., Chen, Z., et al. (2025). Geoanalystbench: a GeoAI benchmark for assessing large language models for spatial analysis workflow and code generation. *Trans. GIS* 29:e70135. doi: 10.1111/tgis.70135

Zuiderwijk, A., Chen, Y.-C., and Salem, F. (2021). Implications of the use of artificial intelligence in public governance: a systematic literature review and a research agenda. *Gov. Inf. Q.* 38:101577. doi: 10.1016/j.giq.2021.101577