GenAI: Substitute or Complement in Evidence Synthesis

A discrete choice experiment with Italian policymakers*

Martina Bazzoli[†] Sergiu Burlacu[†,‡] Dominique Cappelletti[†]
Sonia Marzadro[†] Iunio Quarto Russo[†] Alessio Tomelleri[†] Mirco Tonin[†,§]

April, 2026

Abstract

Credible evidence on the causal effects of public policies is increasingly produced by the academic community, and evidence-based policymaking is increasingly demanded by governments, yet the production of high-quality evidence syntheses remains costly, slow, and often poorly aligned with policymakers' time constraints. Recent advances in generative artificial intelligence (GenAI) promise efficiency gains in evidence synthesis, but their adoption in policy contexts depends on whether policymakers trust and value AI-generated evidence. This paper studies policymakers' preferences for the use of GenAI in producing scientific evidence syntheses for policy design. We implement a discrete choice experiment with 813 Italian policymakers, eliciting trade-offs over synthesis production method (expert-produced, AI-generated, or hybrid AI-with-expert-review), cost, and delivery time. In a between-subject design, we additionally vary whether the AI model is described as general-purpose or specifically validated for evidence synthesis. We find a strong aversion to fully automated AI-produced synthesis, even if it brings efficiency gains. By contrast, the hybrid AI–expert approach is systematically preferred over both expert and AI approaches, conditional on cost and delivery time. Explicit validation of the AI model significantly increases acceptance of AI-assisted synthesis, reducing resistance to automation and increasing preferences for hybrid production. Heterogeneity analyses show that these effects vary with attitudes toward AI and science but not with gender or institutional role. Overall, the results indicate that policymakers value AI primarily as a complement to human expertise rather than as a substitute, highlighting the importance of task-specific validation and expert oversight for the adoption of AI in evidence-based policymaking.

JEL Codes: H11, H83, C90, D73, O33

Keywords: evidence-synthesis, policymakers, AI, DCE

*Data availability: The data and replication code underlying this study will be made publicly available before acceptance. We thank Enrico Rettore, Fabio Cevenini, Clementina Crocé, Matteo Ploner, and Alessandro Tondini for their valuable feedback. We are very grateful to the time dedicated by the policy officials who responded to our request to participate in the study. The research was made possible thanks to a grant from the GRINS (Growing Resilient, INclusive and Sustainable) Partnership, code PE00000018, under Investment 1.3 of the National Recovery and Resilience Plan (PNRR), financed by the European Union - NextGenerationEU, within Mission 4 "Education and Research" - Component 2 "From Research to Enterprise" of the Italian Ministry of University and Research (MUR).

[†]Research Institute for the Evaluation of Public Policies, Bruno Kessler Foundation (FBK-IRVAPP), Via Santa Croce 77, 38122 Trento, Italy

[‡]*Corresponding author;* e-mail: sburlacu@irvapp.it

[§]Free University of Bozen-Bolzano

1 Introduction

Evidence-based policymaking has emerged as a fundamental principle of better governance, built on the idea that government policies ought to be informed by the most reliable and robust research findings available, to maximise the effectiveness and efficiency of public programmes and interventions. On one hand, there is a growing demand for evidence, which has become an institutional priority, with frameworks such as the European Commission's Better Regulation Guidelines or the bipartisan U.S. Foundations for Evidence-Based Policymaking Act of 2018 explicitly requiring impact assessments grounded in scientific evidence, while international organisations like the OECD or the World Bank (see, for instance, the Knowledge for Change Program) have made evidence-informed policymaking a cornerstone of good governance standards (OECD, 2020). On the supply side, the credibility revolution has transformed microeconomics over the past twenty years (Angrist and Pischke, 2010), and social sciences more generally, resulting in a substantial amount of reliable empirical findings on the causal effects of public policies (Currie et al., 2020; Goldsmith-Pinkham, 2024).

In spite of this, policymakers face significant barriers in accessing and utilising this wealth of scientific evidence. Technical expertise limitations, time constraints, access barriers to academic publications, and language barriers constitute major deterrents to incorporating scientific evidence into policy design processes. Research documents persistent challenges, including low statistical literacy, limited understanding of causal research designs, susceptibility to cognitive biases in interpreting evidence, and organisational factors such as transaction costs that influence whether evidence translates into action (Banuri et al., 2019; Briscese and List, 2024; DellaVigna et al., 2024; Hill and Briggs, 2020; Oliver et al., 2014; Pereira et al., 2024; Vivalt and Coville, 2023).

Given these constraints, evidence syntheses—including systematic reviews, rapid reviews, and policy briefs—serve as critical intermediaries between scientific research and policy practice, consolidating vast bodies of research and translating complex findings into accessible formats that facilitate evidence-informed decision-making. Well-designed evidence syntheses can significantly increase policymakers' willingness to adopt evidence-based interventions (Hjort et al., 2021; Toma and Bell, 2024), though engagement remains limited without targeted strategies to demonstrate immediate relevance (Hawkins et al., 2024). A critical challenge is that producing high-quality evidence syntheses requires substantial time and effort from qualified researchers, while policymakers typically require rapid access to evidence to inform time-sensitive decisions—and research suggests that timeliness matters significantly for policy influence (Rao, 2024).

AI can potentially support multiple stages of evidence synthesis tasks, including literature screening to identify relevant studies, data extraction from scientific papers, summarisation of research findings, and drafting of syntheses and policy-relevant summaries. Some scholars suggest that as scientific publication accelerates, AI may increasingly substitute for traditional knowledge brokers in core evidence synthesis functions—identifying relevant research, synthesising findings,

and tailoring summaries for specific policy contexts—fundamentally reshaping the intermediary role between research and policy (Bortolus et al., 2025). This potential has not gone unnoticed by policymakers and practitioners, who are already turning to generative AI for scientific literature reviews, often without formal institutional approval or guidance. A recent survey of 576 public managers across seven EU Member States reveals that generative AI is rapidly becoming a "silent partner" in public administration, with 30% already employing these tools in their daily work and an additional 44% intending to adopt them. Notably, scientific literature reviews emerge as one of the most frequent activities performed by public managers with generative AI: among active users, 72% employ the technology for this specific purpose with varying degrees of frequency - 9% of these users perform this task on a daily basis, 23% several times per week, 26% once a week, and 13% once a month (Tangi et al., 2026).

This widespread informal adoption raises concerns, as standard generative AI models present critical limitations for scientific literature analysis: they may hallucinate information, provide incomplete or imprecise answers, and lack direct access to primary sources and the most recent peer-reviewed scientific publications, making their outputs unreliable for high-stakes policy decisions. Empirical evidence confirms these concerns. A recent comprehensive evaluation of ChatGPT's capabilities in performing systematic literature reviews highlighted significant tension between efficiency and reliability. While the technology demonstrated efficiency gains, with reported workload reductions averaging 60–65% and time savings ranging from 40% to 90%, accuracy proved highly variable across tasks and domains: the model performs well in domains with standardised terminology, such as medical research, but performance drops in disciplines like social sciences or public health, where contextual interpretation is paramount. Most critically, hallucination rates reached 91% in certain tasks, with the model capable of generating fabricated references, misquoted findings, and entirely invented evidence—such as non-existent randomised controlled trials—that appeared plausible to the untrained eye (Adel and Alani, 2025). Beyond hallucinations and accuracy problems, AI-generated evidence syntheses also risk perpetuating systematic biases. For example, it has been found that popular AI tools, including ChatGPT and Microsoft Copilot, consistently generate policy recommendations that overlook gender-specific perspectives, centring male experiences, unless users explicitly specify demographic considerations in their queries. There is the risk that AI may mirror and exacerbate the very discriminatory patterns and structural inequalities that evidence-based policy seeks to address, potentially codifying and legitimising them within formal decision-making processes (Voutyrakou and Skordoulis, 2025).

Beyond these technical and normative problems, the use of generative AI for evidence synthesis raises fundamental epistemic concerns about the nature of knowledge authority in policymaking. Cong-Lem (2025) argues that generative AI has disrupted traditional paradigms of evidence-informed policy by challenging foundational assumptions about what constitutes credible evidence. While conventional evidence synthesis relies on transparent, methodologically rigorous processes with clear provenance and explicit evidence chains, generative AI produces authoritative-

sounding outputs that obscure their underlying logic and sources. This transformation shifts users from consumers of evidence to 'de facto evaluators' who must navigate probabilistic patterns rather than verifiable facts, creating new epistemic risks.

These converging challenges—technical unreliability, systematic bias, epistemic opacity, and informal, unregulated adoption—make understanding policymaker preferences for AI-assisted evidence synthesis both urgent and essential. Promoting evidence-based decision-making remains a cornerstone of effective governance, yet the contemporary challenge has shifted from simply ensuring the use of scientific evidence to ensuring its fair and high-quality use. Policymakers must balance the appeal of efficiency gains against fundamental requirements for transparency, accountability, and equity, as public decisions directly impact citizen welfare in ways that demand not just speed but legitimacy and trustworthiness. The limitations identified above are particularly consequential in this context: policy decisions affect citizens' rights, access to services, and life opportunities, creating non-negotiable standards for accuracy and fairness that cannot be compromised for computational convenience. While the shadow use documented by Tangi et al. (2026) demonstrates that demand for AI-assisted evidence tools exists among public servants, and the efficiency gains documented by Adel and Alani (2025) suggest compelling reasons for this adoption, we lack fundamental understanding of what policymakers actually value when considering such tools for evidence synthesis. Do they prioritise speed over accuracy? Would they accept AI-generated syntheses if sufficiently validated, or do they inherently prefer human expertise? How much are they willing to trade off in terms of cost and time for different production methods and quality assurances? Without systematic evidence on these preferences, efforts to develop responsible AI-assisted evidence synthesis services—whether through validation frameworks, hybrid human-AI models, or enhanced transparency mechanisms—risk being misaligned with policymaker needs and values. Moreover, understanding these preferences is critical for informing AI regulation and governance frameworks that must balance innovation with safeguards appropriate to high-stakes policy contexts. To date, to the best of our knowledge, no research has empirically investigated policymakers' preferences for the key attributes that define evidence synthesis services. Understanding these preferences is essential not only for designing services that policymakers will actually use, but also for ensuring that the inevitable integration of AI into evidence-based policymaking occurs in ways that preserve rather than undermine the epistemic standards and equity commitments that evidence-informed governance requires.

To address this gap, we conducted a large-scale survey experiment embedding a discrete choice experiment (DCE) with Italian policymakers and public officials. The study was designed to elicit preferences for alternative models of producing policy-relevant scientific evidence syntheses under constraints. Respondents were asked to imagine being responsible for developing a policy proposal within a limited time frame on a highly relevant issue in their area of competence, and with a budget to commission a scientific evidence synthesis to support policy proposal development. In the DCE, they were asked to choose between competing evidence syntheses that differed along

three dimensions: the production method (expert-produced, AI-generated, or hybrid AI-with-expert-review), delivery time, and cost. All syntheses were described as being produced from the same underlying corpus of scientific evidence, isolating perceptions of the production process rather than differences in content. In a between-subject design, we additionally varied whether the AI model is described as general-purpose or specifically validated for evidence synthesis. This design allows us to examine not only preferences over AI versus human expertise, but also the role of validation and task-specific credibility in shaping acceptance of AI-generated evidence.

The results reveal three main findings. First, policymakers exhibit a strong aversion to fully automated AI-generated evidence syntheses. Even when AI-only production offers dramatic reductions in cost and delivery time, it is strongly disfavoured relative to syntheses produced by human experts. This pattern suggests that efficiency gains alone are insufficient to compensate for concerns related to reliability, accountability, and epistemic authority in the policy context.

Second, policymakers prefer the hybrid production approach in which generative AI is used to produce an initial synthesis that is subsequently reviewed and revised by human experts. Conditional on cost and delivery time, this hybrid approach dominates both the expert-only status quo (at least until recently) and the fully automated AI option. This preference suggests that policymakers do not view AI merely as a cost-saving substitute for human labour, but rather as a complementary input that can improve expert productivity while preserving human oversight and expertise.

Third, explicit validation of the AI model plays a key role in shaping acceptance. Describing the AI system as specifically tested and validated for evidence synthesis substantially reduces resistance to automation and increases the relative preference of both AI-only and hybrid approaches. Validation narrows, but does not eliminate, the preference gap between AI-only and expert-only production, and increases the premium placed on the hybrid approach. This finding underscores the importance of task-specific credibility signals in contexts where policymakers cannot directly assess the technical quality of AI outputs.

We further explore heterogeneity in preferences across individual characteristics. The strongest differences emerge along policymakers' attitudes towards AI: those with more positive views are more willing to accept AI-assisted synthesis, particularly under validation, while those with less positive attitudes show strong resistance to automation. Attitudes towards science reveal a more nuanced pattern: respondents with more positive attitudes do not differ significantly in their aversion to fully automated AI, but become more favourable to the hybrid model when the AI system is presented as validated for evidence synthesis. This suggests that they value AI mainly as a complement to expert judgement, rather than as a substitute for it. Finally, preferences are generally stable across political versus administrative roles and gender.

To summarise, the results provide systematic evidence that policymakers value AI primarily as a complement to human expertise rather than as a substitute. While efficiency matters, human oversight remains key, whereas task-specific validation is strongly valued by policymakers

in order to increase acceptance of AI in knowledge-intensive public sector tasks.

Our study makes several contributions to the literature on artificial intelligence in public administration, evidence synthesis, and evidence-based policymaking. First, to the best of our knowledge, this is the first study of policymakers' preferences and attitudes towards the use of AI in public administration specifically focusing on AI as a tool for producing policy-relevant scientific evidence. While existing research has examined algorithmic decision-making and automation within public agencies (e.g., Levy et al., 2021; Alon-Barkat and Busuioc, 2023; Hillo et al., 2026), little is known about how policymakers themselves evaluate AI when it supports upstream knowledge production rather than final decisions. Our findings provide quantitative support for a hybrid AI-expert model over both pure human and pure AI approaches, addressing a critical design question for evidence synthesis services. Importantly, this preference aligns with emerging best practices in the literature. For example, Adel and Alani (2025) propose a Systematic Research Processing Framework (SRPF) that advocates for precisely such hybrid collaboration, where AI augments efficiency in structured tasks like initial screening while human experts retain responsibility for contextual validation, ethical oversight, and interpretive synthesis. Thus, while research from computer science suggests – from the supply side – that the optimal integration of AI in evidence synthesis lies not in full automation but in strategically designed human-AI collaboration that leverages the comparative advantages of each, our empirical findings complement this from the demand side, showing that policymakers value precisely such human-AI collaboration.

This preference for hybrid models resonates with broader evidence on citizen attitudes towards AI in government: research on public acceptance of algorithmic decision-making consistently finds that citizens view purely automated systems as illegitimate and undemocratic, while hybrid arrangements where humans remain in control are considered acceptable (**?**). The presence of human oversight functions as a crucial legitimacy signal, with studies showing that "human-in-the-loop" assurances significantly increase public trust in AI-assisted decisions, particularly in sensitive policy domains (Aoki, 2021). Thus, policymaker preferences for hybrid evidence synthesis not only align with technical best practices but also with democratic expectations: maintaining human involvement preserves both the epistemic quality that experts provide and the democratic legitimacy that citizens demand. Our finding of a strong aversion to fully automated synthesis also aligns with recent evidence from the private sector, where managers' acceptance of AI technology has been shown to decline sharply once functionality crosses a 'critical boundary' into autonomous implementation without prior human control (Gieselmann et al., 2025). Together, these results reinforce a cross-sectoral preference for human oversight, suggesting that decision-makers in both public and private domains value AI primarily as a productivity-enhancing complement rather than a substitute for expert judgment in high-stakes contexts.

Second, we provide novel evidence on the critical role of AI validation in shaping policymakers' acceptance. By experimentally manipulating whether AI is described as a general-purpose model or as specifically validated for scientific evidence synthesis, we isolate the effect of validation

signals on stated preferences. Our findings demonstrate that explicit validation substantially increases policymakers' willingness to accept AI-assisted evidence synthesis, narrowing the preference gap between AI-only and expert-only approaches. This finding is particularly significant because it directly addresses a primary barrier to AI adoption in government consistently identified across empirical studies: lack of confidence and trust in the technology (Tangi et al., 2026; Wirtz et al., 2019). This barrier has been conceptualised as "technoskepticism"—scepticism rooted in AI's opacity and the difficulty of independently verifying the quality and reliability of algorithmically-generated knowledge claims (Newman and Mintrom, 2023).

Our results suggest that task-specific validation can function as a credibility signal that mitigates this scepticism, providing policymakers with procedural assurance about AI reliability even when they cannot directly assess the technical quality of outputs. This aligns with broader insights from information theory showing that decision-makers under time constraints and information asymmetries rely on signals about how information is produced to infer trustworthiness (Lupia and McCubbins, 1998). Moreover, our finding has direct policy relevance: the distinction between general-purpose AI and task-validated AI is increasingly central to governance and regulatory frameworks, which explicitly emphasise validation-for-use as a prerequisite for deploying AI in high-stakes analytical contexts (European Union, 2024; National Institute of Standards and Technology, 2023). By demonstrating that policymakers themselves differentiate between validated and unvalidated AI—and that this differentiation materially affects their willingness to use AI for evidence synthesis—our research provides empirical support for regulatory approaches that mandate validation and transparency as mechanisms to enhance AI trustworthiness and appropriate reliance in policy contexts.

Third, our findings contribute evidence on policymakers' preferences that speaks to concerns shared across national contexts. Evidence-based policymaking is increasingly recognised as a core component of good governance across countries and institutional settings (OECD, 2020), such that the value of efficient and trustworthy evidence synthesis, and policymakers' preferences for how it is produced, is increasingly relevant across governmental contexts worldwide. The behavioural patterns we document: aversion to full automation, preference for hybrid models, and the role of validation, converge with a growing body of related findings across countries and institutional contexts (Gieselmann et al., 2025; Alon-Barkat and Busuioc, 2023; Hillo et al., 2026).

Fourth, we contribute methodologically by employing a discrete choice experiment to investigate policymakers' preferences for evidence synthesis production methods. While DCEs have been increasingly used to study how policymakers evaluate research evidence—including preferences for research design characteristics (Nakajima, 2021), local versus external expertise (Vivalt et al., 2025), evidence quality signals and author credibility (Xu et al., 2024), and evidence synthesis methodological shortcuts (Speckemeier et al., 2021)—to our knowledge, this is the first application specifically examining preferences for AI versus human expert production of evidence syntheses. This distinction is important: previous DCE studies have focused on attributes of the evidence

itself (sample size, research design, contextual relevance, author affiliation) or characteristics of human-produced evidence, whereas we investigate preferences for the production process when AI is introduced as a potential substitute or complement to human expertise. Discrete choice experiments offer distinct advantages over traditional survey methods by forcing respondents to make realistic trade-offs between competing attributes rather than rating each dimension in isolation. This approach enables us to quantify not just which attributes matter, but how much they matter relative to each other. Moreover, our experimental design incorporates a between-subjects treatment manipulation, randomly assigning respondents to conditions where AI is described as either general-purpose or specifically validated for scientific synthesis. This experimental component allows us to causally identify the effect of validation signals on stated preferences, isolating this mechanism from potential confounding factors present in observational studies.

2 Research design

This section describes the research design, beginning with the construction of the sample pool of policymakers, the survey instrument, and the discrete choice experiment. Before going into details, we provide a brief overview of the study context, which shaped the research design and research questions, driven and motivated by practical considerations related to the development of a tool intended to support policymakers in evidence-based decision-making.

The study was part of the *PolicyAId* project (*Policy Artificial Intelligence Digest: AI-powered platform for evidence-based policy decision*), funded through the Italian National Recovery and Resilience Plan within the GRINS Partnership, aimed at developing a demand-driven and cost-efficient tool for Italian policymakers and public officials to facilitate evidence-based decision-making. The project was completed in late 2025 and is currently integrated into AMELIA, a comprehensive digital research infrastructure developed by the GRINS (Growing Resilient, INclusive and Sustainable) Foundation to support evidence-informed decision-making across the full policy cycle. The current study was conducted during the development stages of the project. Its primary goals were to understand the needs and barriers associated with evidence-based policy-making, assess attitudes towards AI and science, and estimate preferences for different approaches to evidence-synthesis production. The study was pre-registered on the American Economic Association's registry for randomised controlled trials (ID number AEARCTR-0016366).

2.1 Sample

The survey sample was constructed by compiling a comprehensive database of email contacts of Italian policymakers. This strategy was adopted to ensure broad institutional coverage and to reduce potential selection biases associated with network-based recruitment methods, such as snowball sampling, which may limit participation to individuals already connected to academic or research networks. The target population included political and administrative actors with decision-making

roles across multiple levels of government and public administration, ranging from central institutions (the national Parliament) to sub-national and local authorities (Regions, Autonomous Provinces, Metropolitan Cities, and Municipalities).

For the Parliament, Regions, Metropolitan Cities, and Municipalities with more than 100,000 inhabitants, named email addresses of policymakers were collected through systematic web scraping of official institutional websites. Web scraping followed a semi-automated, adaptive procedure tailored to the heterogeneous structure of institutional websites. Pages containing relevant information were first identified manually; these typically consisted either of individual profile pages or of aggregated pages listing the members of a department, office, service, or administrative unit. Depending on the structure of the webpage, data were then extracted either through custom Python scripts or by submitting the HTML body content to a large language model (GPT) to obtain structured, tabular outputs. In all cases, human oversight was retained throughout the process. Specifically, manual validation was used to define and select which roles could be classified as policymakers: all elected officials were included among political roles, whereas only administrative positions involving decision-making authority or managerial responsibility (e.g., heads of units, managers, directors) were retained among administrative roles. GPT was additionally employed to infer the apparent gender of individuals from their first names when this information could not be unambiguously derived from role titles; this variable was used exclusively for the grammatical personalisation of invitation letters.

For the remaining entities—namely, Municipalities with fewer than 100,000 inhabitants—generic (non-named) email addresses of the municipality or of relevant offices were obtained from a commercially available online mailing list. In four Municipalities with more than 100,000 inhabitants, named email addresses of policymakers were not available on institutional websites; in these cases, a generic, non-named email address was used instead.

Overall, the dataset comprised 5,209 named email addresses, with a varying proportion of political and administrative roles, and 7,828 non-named email addresses, for a total of 13,037 contacts. Invitation emails were sent between 22 July and 8 August 2025, from Monday to Friday between 8:00 AM and 6:00 PM, with a maximum of 900 emails sent per day. Named email addresses received a personalised invitation email with an attached invitation letter addressed directly to the recipient, whereas non-named municipal addresses received an email requesting internal forwarding to members of the municipal council and executive, as well as to senior managers and officials, with an invitation letter addressed generically to the institution. Both the email messages and the attached letters included a link to access the online survey.

2.2 Survey

The survey started with a detailed background questionnaire. The first section captures the professional and institutional profile of the respondent, to distinguish between political and adminis-

trative roles and identify their specific level of authority, such as managerial positions or executive versus legislative functions. It also asks for the territorial jurisdiction and geographical macro-area. To contextualise the level of involvement in the policy design, the survey mapped professional activities performed in the preceding year, including tasks such as studying policy solutions, monitoring implementation, or conducting impact evaluations, alongside their specific macro-sectors of expertise. This was followed by standard socio-demographic information, including gender, age, and educational attainment.

The final two sections of the questionnaire explored the respondents' current practices and attitudes regarding scientific evidence / evidence-based policymaking and artificial intelligence. To balance out any potential priming effects, in particular regarding attitudes towards science and AI, on the experimental choices, the order of these two sections was randomised at the individual level. In particular, one section measures the frequency with which policymakers consult scientific literature and identifies barriers that may impede the use of evidence in policymaking. After indicating a topic for which they would be interested in receiving a scientific synthesis (e.g., improving recycling rates, addressing declining birth rates, regulating short-term rentals), respondents completed a scale assessing attitudes towards the role of science in general and specifically within the policy process (henceforth "Attitudes towards Science" for brevity), using three items from the six-item battery developed by Cologna et al. (2025) to measure normative perceptions of science in policymaking (e.g., "Scientists should work closely with politicians to integrate scientific results into policy-making."). Our data confirm good internal consistency for this scale (Cronbach's $\alpha = 0.75$). The other section quantified the respondents' frequency of generative AI use and their professional attitudes towards AI through the three-item ATTARI scale from Gnambs et al. (2025), specifically the version on attitudes towards AI in the workplace. Our data confirm good internal consistency also for this scale (Cronbach's $\alpha = 0.81$).

2.3 Discrete choice experiment

To identify policymakers' preferences for different characteristics of scientific evidence, we implemented a discrete choice experiment (DCE). A DCE is an experimental design used to elicit individual preferences by asking respondents to state their choice over a set of hypothetical alternatives. Each alternative is defined by a combination of attributes at varying levels; by observing how choices change as levels vary, we can estimate the relative importance and trade-offs between these characteristics.

The framing of the experiment was tailored to the professional context of policymakers and public servants. Respondents were instructed to imagine being tasked with preparing a proposal for a new policy or programme on a highly relevant topic in their area of expertise, to be presented to a relevant committee within three months. They were informed that a synthesis of scientific studies evaluating the effects of public policies or programmes on similar topics would be made available to support this work. Under this scenario, participants were asked to choose between

different types of evidence syntheses which differed in how they were produced, delivery time and cost. The vignette was designed to create a realistic scenario suitable for our context, drawing on insights from the qualitative interviews conducted with 25 policymakers prior to the survey, which indicated that policymakers are typically required to design policy proposals under time constraints.

Each respondent completed eight choice tasks. In each task, two profiles of evidence syntheses were displayed side-by-side, and respondents selected the profile they considered more useful for informing the policy proposal. Each profile had three attributes. To control for potential ordering effects, the order of these three attributes was randomised across respondents. However, this remained constant across all eight choice tasks for any given individual, to avoid excessive cognitive burden, in line with the recommendations in the literature (Johnston et al., 2017). We kept the number of attribute values low, resulting in 45 possible profiles, which, given the sample size, should ensure high balance across attribute levels. Given the expected sample size, we opted for a fully randomised design with constraints to ensure realism but with overlap across options, which we detail below.

The three attributes - synthesis production type, delivery time, and cost - and their levels are presented in Table 1. The first attribute is the one of primary interest for the study, describing the approach used in producing the syntheses. We impose the constraint that, in a given choice set, the levels of this attribute are different across alternatives. Importantly, we stressed and represented also visually the fact that all three attribute levels start from the same corpus of scientific evidence, to avoid choices being driven by concerns regarding access of AI models to licensed articles or poor AI selection of relevant literature. The first level reflects the status quo, that is, an evidence synthesis produced by experts (i.e., academics or consultants with experience on the topic), with longer delivery time and higher costs. The second level describes a process fully automated with a generative AI model, with shorter delivery time and lower cost. The last level describes a hybrid approach, where a generative AI model produces a draft of the synthesis, which is then revised by experts, with delivery time and cost levels standing in between the previous two. To limit extrapolation, despite the constraints imposed for realism, we also allowed overlap across production types' cost and delivery time, as detailed in Table 1 below.

Table 1: Discrete Choice Experiment: Attributes and Levels

Production Method	Delivery Time (days)[a]	Cost (€)
By experts	14, 30, 60	1000, 1500, 2000, 3000, 4000
By a GenAI model	1, 7, 14	250, 500, 1000, 1500, 2000
By a GenAI model followed by revision from experts	7, 14, 30	500, 1000, 1500, 2000, 3000

Notes: [a] Delivery time attribute levels were expressed in more commonly occurring temporal references (e.g. one day, one week, one month). Double-underlined attribute levels are shared across all three synthesis production types; underlined attribute levels are shared among two production types. The vignette explicitly stated and visually presented that all three production approaches start from the same corpus of scientific evidence. The actual description of the attributes and levels in the experimental instructions was more detailed to ensure comprehension; for full details, see dedicated section in the Online Appendix.

We did not directly include an opt-out option in the choice cards because participants often use it to circumvent the cognitive burden of choosing between the presented alternatives (Dhar and Simonson, 2003).

2.4 Between-subject Treatment: model validation

Motivated by the ongoing policy and scholarly debate on the reliability, trustworthiness, and transparency of generative AI systems, we embedded a between-subject treatment that varied whether the generative AI model producing the evidence synthesis was described as a general-purpose model or as a model validated for the specific task of evidence synthesis. Concerns about large language models, their tendency to hallucinate, and the difficulty of assessing the reliability of their outputs have been widely discussed in both academic and policy circles (Floridi et al., 2018; OECD, 2023; European Commission, 2024). In response, recent regulatory and governance frameworks increasingly emphasise model validation, benchmarking, and task-specific evaluation as key mechanisms to enhance trust in AI-enabled systems.

Because we did not expect all policymakers to be familiar with technical AI terminology or distinctions commonly used in the machine learning literature, we used simple and intuitive labels—"generalist" and "validated"—for the two treatment arms, and provided short, accessible explanations of what these terms imply in practice. The generalist AI model was described as a "generative AI for general purposes, trained on a wide variety of texts available online, but not specifically tested in the synthesis of scientific studies." By contrast, the validated AI model was explained as a "generative AI specifically tested in the synthesis of scientific studies, by comparing the results it produces with those generated by human experts."

This design choice reflects a broader distinction in the literature between general-purpose AI systems and AI systems that are evaluated and validated for specific tasks. In high-stakes or specialised applications, governance and auditing frameworks increasingly treat task-specific evaluation as a prerequisite for establishing credibility and for calibrating users' reliance on AI-generated outputs (Raji et al., 2020; Bommasani, 2021). In the context of evidence synthesis, validation against expert benchmarks is increasingly viewed as a necessary condition for the responsible use of AI tools, particularly when outputs are intended to inform policy decisions (Flemyng et al., 2025). By experimentally varying whether such validation is made explicit, our design allows us to assess whether and to what extent policymakers value task-specific validation as a signal of credibility and reliability when evaluating AI-generated evidence.

2.5 Empirical specification

We estimate policymakers' preferences using standard discrete choice models grounded in random utility theory (McFadden, 2001). In each of the eight choice tasks, $t = 1, \ldots, 8$, respondent i chooses between two evidence-synthesis alternatives, indexed by $j \in \{1, 2\}$. Utility is specified as

the sum of a systematic component, determined by the experimentally varied attributes, and an idiosyncratic error term:

$$U_{ijt} = V_{ijt} + \varepsilon_{ijt}. \tag{1}$$

Our baseline model is a conditional logit model, equivalent to a multinomial logit model (MNL) for alternative-specific attributes, in which the utility that the respondent derives from choosing an evidence synthesis alternative in a given choice task is a linear function of the attributes describing the synthesis—production method, delivery time, and cost, plus an idiosyncratic error term. The systematic component of utility is specified as:

$$V_{ijt} = \beta_1 AI_{ijt} + \beta_2 Hybrid_{ijt} + \beta_3 Cost_{ijt} + \beta_4 Time_{ijt}. \tag{2}$$

AI_{ijt} indicates whether the synthesis is produced automatically by a generative AI model, while $Hybrid_{ijt}$ indicates that the synthesis is first produced by a generative AI model and then reviewed by experts. The omitted category is a synthesis produced entirely by experts. The coefficients β_1 and β_2 capture the relative utility of fully automated and hybrid production compared with the expert-only benchmark, conditional on cost and delivery time. We estimate the model by maximum likelihood and cluster standard errors at the respondent level to account for repeated choices. To facilitate interpretation, willingness-to-pay measures are computed by normalising attribute coefficients by the cost coefficient and reported in Table A7 in the Online Appendix.

To estimate the effect of task-specific AI validation, we exploit the between-subject experimental design. Respondents were randomly assigned either to a condition in which the AI model was described as general-purpose or to a condition in which it was described as validated for evidence synthesis. We estimate separate models by treatment arm for exposition and formally test differences in the production-method coefficients using the following pooled interaction specification:

$$\begin{aligned} V_{ijt} = {} & \beta_1 AI_{ijt} + \beta_2 Hybrid_{ijt} + \beta_3 Cost_{ijt} + \beta_4 Time_{ijt} \\ & + \delta_1 (AI_{ijt} \times Validated_i) + \delta_2 (Hybrid_{ijt} \times Validated_i). \end{aligned} \tag{3}$$

Here, δ_1 and δ_2 measure whether describing the AI model as validated for evidence synthesis changes respondents' valuation of fully automated AI production and hybrid AI–expert production. Since treatment assignment is constant within respondents and does not vary across alternatives within a choice task, its main effect is not separately identified in this choice design.

To examine observed heterogeneity, we interact the production-method indicators with respondent-level characteristics, including attitudes towards AI, attitudes towards science, gender, and institutional role. These characteristics do not enter the utility function directly, because they do not vary across alternatives within a choice task and are therefore identified only through

interactions with alternative-specific attributes.

We finally relax the fixed-coefficient assumption of the multinomial logit model by estimating mixed logit models. To account for unobserved heterogeneity in preferences, these specifications allow the coefficients on the synthesis production methods, cost, and delivery time to vary randomly across respondents, with random parameters assumed to follow normal distributions.

3 Data: Descriptive Statistics, Balance and Attrition Check

A total of 898 respondents started the discrete choice experiment. Of these, 38 respondents (4.2%) exited during the instructional stage, while an additional 47 respondents (5.2%) dropped out while performing the conjoint tasks (Table A3). We define as our analytic sample the 813 respondents who completed all eight tasks. Although overall attrition is low (9.4%), in the Online Appendix we assess (i) whether attrition rates differ by between-subject treatment arms (Table A4), and (ii) whether attrition is correlated with respondent characteristics (Table A5). We do not find either to be a major concern for experimental validity: although AI attitudes are negatively correlated with attrition in the treatment group, a joint model including all treatment arm interactions with respondent characteristics fails to reject the null hypothesis that all interaction terms are zero ($p = .906$). In addition, Table A6 reports balance checks by treatment arm for the analytic sample and finds no relevant differences.

Table 2 reports the survey descriptive statistics for the analytic sample, i.e., role-specific statistics and demographic characteristics. The sample is balanced between respondents holding political roles (50.8%) and those in administrative roles (49.2%). Among respondents with a political role, the vast majority serve in representative or executive institutions: 65.1% are members of a legislative or assembly body, while 32.4% hold an executive office. Only a very small share (1.2%) reports an active position in a political party, movement, or labour union as their primary political role. The sub-sample of respondents with an administrative role is composed primarily of officials with managerial or director-level positions (69.2%), consistent with our sampling strategy targeting individuals expected to have decision-making authority.

Respondents operate across a wide range of territorial jurisdictions, but the distribution is weighted towards local governments, 44.8% serving in municipalities with fewer than 20,000 inhabitants and 28.4% in municipalities above that threshold. A further 8.4% operate at the provincial or metropolitan city level, and a notable 17.6% hold positions within regional administrations. Less than 1% report working at the national or European level. In terms of geographical distribution, approximately 76% of respondents are located in the North-East or North-West, 14.9% in Central Italy, and 9.3% in the South or Islands.

Turning to the demographic characteristics, the gender representation is roughly 59% male and 41% female respondents. The age distribution is skewed towards mid-career and late-

career public officials: only 10.5% of respondents are 35 or younger, while the largest shares fall between 36 and 55 (44.3%) and between 56 and 65 (35.1%). Nearly 10% are above 65. The overall age pattern appears to be consistent with the tenure structures and seniority requirements typical of public sector leadership roles. Tables A1 and A2 in the Online Appendix compare sample characteristics with registry data for local and regional politicians and administrative managers, respectively. Gender and age distributions are closely aligned with the reference populations, while politicians in our sample display a higher level of education, though education data are missing for almost a quarter of registry entries.

Table 2: Descriptive statistics of the sample ($N = 813$)

	%
Political Role[a]	50.8
Executive body member	32.4
Legislative body/assembly member	65.1
Party, movement, or union official	1.2
Administrative Role[a]	49.2
Manager / Director	69.2
Officer / Civil servant	24.5
Other employee	5.5
Consultant	0.8
Territorial jurisdiction	
Municipality (< 2,000 inhabitants)	12.1
Municipality (2,000 – 20,000 inhabitants)	32.7
Municipality (> 20,000 inhabitants)	28.4
Province / Metropolitan City	8.4
Regional	17.6
National or European	0.7
Geographical Area	
North East	46.2
North West	29.6
Center	14.9
South	4.7
Islands	4.6
Gender	
Female	40.6
Male	58.8
Age	
18–35 years	10.5
36–55 years	44.3
56–65 years	35.1
Over 65 years	9.7
Education	
High school or less	23.6
University degree	52.3
Post-graduate degree	23.7

Note: Residual categories (e.g., "Other" or "Prefer not to respond") are not reported in the table, which may cause some sections not to sum to exactly 100%. Geographical classifications follow standard definitions by ISTAT, the national statistical office.
[a] Percentages for indented sub-categories are calculated relative to the total number of respondents within that specific role (Political $n = 413$; Administrative $n = 400$).

Following the initial professional and demographic profiling questions, respondents completed two additional survey sections, in random order to mitigate potential priming effects on subsequent choices in the discrete choice experiment. One section focused on the frequency of use of scientific evidence, perceived barriers to its use, and attitudes towards science. Similarly, another section examined the frequency of use of, and attitudes towards, generative AI. In what follows, we present descriptive statistics from these sections.

Italian policymakers display substantial heterogeneity in the frequency with which they consult scientific evidence, either directly or indirectly through collaborators or external consultants. While approximately 30% of the respondents report consulting scientific evidence often, 39% indicate doing so only sometimes. A non-negligible share reports very limited use, with 21% consulting it rarely and 6% never doing so. Despite this variation in use, perceptions of the usefulness of scientific studies are positive. A majority of respondents report finding scientific evidence quite useful (54%) or very useful (38%). Respondents were also asked to identify, from a precompiled list, the main barriers preventing them from making more frequent use of scientific evidence. The most commonly cited obstacles are lack of time (53%) and lack of dedicated personnel (41%). These are followed by difficulties in translating scientific evidence into actionable policy proposals (37%), difficulties in identifying relevant studies (36%), and the lack of syntheses of scientific evidence (33%). The high incidence of the latter two barriers, in particular, points to a demand for timely and usable evidence syntheses tailored to policy-making needs. Respondents display generally positive attitudes towards science: the normalised score distribution is shifted towards the theoretical maximum of 1 and has a mean and median of 0.73 and 0.75, respectively.

Heterogeneity is even more pronounced in the reported use of generative AI, for which respondents were provided with examples of widely known existing models. More than half of respondents use AI, either sometimes (32%) or frequently (19%). The rest report using generative AI either rarely (26%) or never (23%). Consistent with these patterns, we observe considerably greater heterogeneity in attitudes towards generative AI than in attitudes towards science. The distribution of the resulting normalised index is centred at 0.58 of its theoretical maximum, indicating more ambivalent attitudes towards AI compared to the generally positive views expressed towards science.

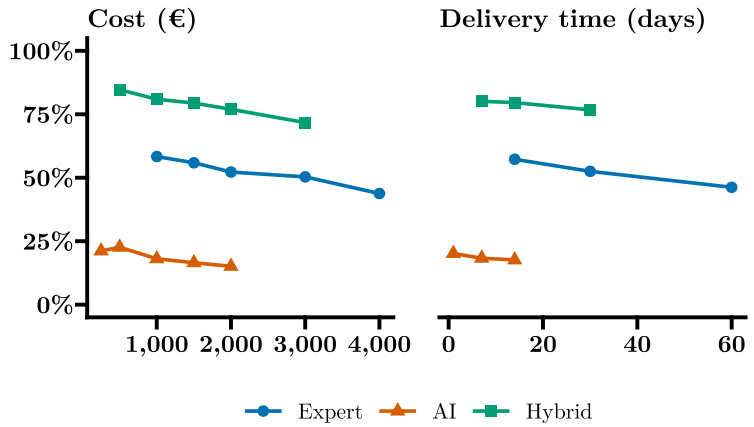
4 Results

Before diving into multinomial logit models and mixed logit models, we present a descriptive analysis of choice frequencies across the three synthesis production types. Across all choice sets and between-subject treatments, the results indicate a stronger preference for the Hybrid production option, selected in 79% of the instances where it was offered. This is somewhat surprising: our ex-ante expectation was that the status quo, i.e., the Expert approach, would be the preferred approach, conditional on costs and delivery time. What we find, however, is that the Expert has a choice share of 52%. Despite the much lower cost and delivery time, the automatic AI production approach was selected only 19% of the time. This strong variation in raw choice shares suggests that while human expertise is valued, the combination of human and artificial intelligence in a hybrid format provides the highest utility to respondents, a finding we will test further by controlling for other attributes in our formal discrete choice models.

To supplement the results presented above, Figure 1 reports choice frequencies for each

synthesis production type disaggregated by cost and delivery time levels. These descriptive plots indicate that choice shares decrease linearly as both cost and delivery time increase, indicating that these two attributes are relevant for policymakers and suggesting no substantial non-linearity or threshold effects. The slopes of these curves show only some minor variation across the three synthesis types. This suggests that while the baseline preference for synthesis type remains dominant, respondents maintain a relatively consistent sensitivity to efficiency and cost constraints across all production formats.

Figure 1: Choice share by synthesis production type, cost, and delivery time



Notes: Expert indicates an evidence synthesis produced by experts (i.e., academics or consultants with experience on the topic). AI indicates a process fully automated with a generative AI model, while Hybrid indicates a synthesis generative AI model, which is then revised by experts. Total $N = 813$ respondents $\times$ 8 choice tasks $\times$ 2 alternatives.

4.1 Main effects

We begin with the results from the multinomial logit (MNL) models. Table 3 reports the estimated marginal utilities for the entire sample in Column (1) and separately by treatment arm in Columns (2) and (3). The reference category for the synthesis production attribute is the status quo, an evidence synthesis produced by experts. While we present treatment-arm split-sample results for exposition, statistical significance of the treatment effects is formally assessed using models with interaction terms between attributes and treatment, reported in Table A8 in the Online Appendix.

The estimates are strongly in line with the descriptive results presented above. The results for the entire sample (Column 1) suggest that policymakers strongly prefer the status quo over an evidence synthesis produced automatically by a generative AI model without human supervision ($\beta = -1.76$, $p < .001$). In addition, the MNL estimates confirm the premium on the hybrid approach. Even when conditioning on cost and delivery time, that is, holding efficiency gains constant, a synthesis produced automatically by a generative AI model but checked and revised by experts is preferred over the status quo ($\beta = 0.49$, $p < .001$). This suggests that policymakers do not merely view AI as a cost-cutting tool, but perceive intrinsic value in an

AI-human approach. This finding aligns with the concept of "algorithm appreciation" when a human-in-the-loop serves as a safety net (Logg et al., 2019), likely mitigating hallucinations and other risks while promising a more comprehensive or unbiased review of the literature than a group of experts might provide. This preference for hybrid models — where AI complements rather than substitutes human expertise — is also consistent with evidence from the Italian public sector showing that public employees generally exhibit positive attitudes towards AI, viewing it favourably as an augmentative tool rather than a replacement for human roles Giraldi et al. (2024).

Table 3: Preferences for Evidence Synthesis (Multinomial Logit)

	Full Sample	Generalist AI	Validated AI
Automatic AI	-1.759***	-2.261***	-1.308***
	(0.092)	(0.139)	(0.121)
Hybrid (AI + Expert)	0.488***	0.269**	0.710***
	(0.067)	(0.093)	(0.097)
Cost (€000)	-0.260***	-0.277***	-0.252***
	(0.029)	(0.042)	(0.040)
Time (weeks)	-0.081***	-0.082***	-0.082***
	(0.013)	(0.019)	(0.019)
Log-Likelihood	-3274.5	-1565.5	-1652.3

Notes: Coefficients estimated via multinomial logit. Clustered standard errors at respondent level in parentheses. The reference category for the synthesis method is the "Expert" produced synthesis status quo. The "Automatic AI" option refers to synthesis produced by a general-purpose model (*Generalist AI*) or a validated model (*Validated AI*). Total $N = 813$ respondents $\times$ 8 choice tasks $\times$ 2 alternatives. Asterisks indicate statistical significance: * $p < .05$, ** $p < .01$, *** $p < .001$.

Comparing Columns (2) and (3) reveals substantial treatment heterogeneity. In the control group (exposed to a general-purpose AI description), the penalty on fully automated synthesis is substantially larger ($\beta = -2.26$) than in the treatment group (validated AI model) ($\beta = -1.31$). Model validation reduces resistance to automation: policymakers become more likely to substitute an expert-produced synthesis with one produced automatically by a GenAI model if the model has been parametrised and evaluated specifically for the task at hand. Trust in and acceptance of AI models to provide support in evidence-based policymaking require assurance of quality for the specific task.

Furthermore, treatment significantly increases the preference for the hybrid approach (β increases from 0.27 in the Control group to 0.71 in the Treated group). This indicates that while openness to fully substituting experts remains limited, though greater when the model is specifically validated for the purpose, policymakers perceive a strong complementarity between AI

and experts. A crucial requirement in interacting with AI tools to enhance productivity is strong domain-specific knowledge, which experts (i.e., researchers) have. Policymakers' choices seem to align with this idea: given that careful experts can presumably spot hallucinations and other errors produced by the model, AI can be an important tool for raising researcher productivity, potentially also reducing biases and increasing the reliability of the research evidence considered. For instance, if experts are more likely to weigh evidence from domains, fields, or journals more familiar to them, they may discard important evidence that AI might instead consider. Thus, having a validated AI model increases not only the relative value of a fully automated AI tool but also that of a hybrid one.

To better put marginal utilities into perspective, we convert them to willingness-to-pay estimates by normalising them against the cost coefficient (Table A7 in the Online Appendix). In the control group, policymakers are willing to pay approximately €8,155 to avoid a fully automated synthesis in favour of experts. In the treatment group, this premium drops to €5,186, reflecting increased trust in the validated AI. Both estimates exceed the maximum price level in the experimental design, consistent with extrapolation driven by a share of respondents who, even at very low relative prices, never choose the automated AI option. The hybrid approach shows a positive premium over the human expert: policymakers are willing to pay an additional €970 in the control group and €2,816 in the treatment group for a synthesis produced through the hybrid method, ceteris paribus. The delivery time WTP is roughly €300 per week.

Also factoring in the delivery time coefficient, we create a realistic policy scenario that accounts for the efficiency gains of AI. We compare an "Expert" profile with high costs and slow delivery (€4,000, 60 days) against an automatic AI profile with minimal costs and delivery (€250, 1 day). Under this scenario, the premium for Expert relative to AI drops to roughly €1,900 in the control group and a negative €-1,300 in the treatment group, where the latter suggests that the efficiency gains of a validated model can lead to a substitution effect. When we apply a similar efficiency scenario to the Hybrid approach (assuming a 50% gain: €2,000 cost, 30 days delivery), the premium for the Hybrid option reaches roughly €4,240 in the control group and €6,210 in the treatment group.

4.2 Heterogeneity in preferences

We explore heterogeneous effects across several dimensions: attitudes towards and use of AI, attitudes towards science (specifically regarding the normative engagement of scientists in policymaking), use of scientific evidence, and the perceived usefulness of scientific evidence. For the analysis presented in the following, we interact the treatment and attribute variables with indicator variables for these dimensions, constructed using median splits for continuous scales and balanced grouping for categorical variables. Additionally, we perform an exploratory heterogeneity analysis by institutional role (political vs. administrative) and gender. We report in the main text only the most relevant dimensions and only discuss the remaining ones.

Table 4: Heterogeneity Analysis by Attitudes toward AI and Science

	Generalist	Validated	Generalist	Validated
Automatic AI	-2.779***	-1.929***	-2.101***	-1.382***
	(0.200)	(0.175)	(0.168)	(0.156)
Hybrid (AI + Expert)	-0.154	0.275*	0.217	0.476***
	(0.132)	(0.132)	(0.119)	(0.126)
AI $\times$ $\mathbb{1}_{\{X>\text{med}(X)\}}$	0.891***	1.127***	-0.396	0.126
	(0.219)	(0.197)	(0.216)	(0.192)
Hybrid $\times$ $\mathbb{1}_{\{X>\text{med}(X)\}}$	0.842***	0.862***	0.106	0.454**
	(0.168)	(0.177)	(0.171)	(0.176)
Cost (€000)	-0.283***	-0.263***	-0.277***	-0.256***
	(0.042)	(0.042)	(0.042)	(0.041)
Time (weeks)	-0.091***	-0.084***	-0.084***	-0.083***
	(0.019)	(0.020)	(0.019)	(0.019)
Log-Likelihood	-1529.4	-1600.4	-1559.1	-1645.0
$X =$	AI attitudes	AI attitudes	Science attitudes	Science attitudes

Notes: Coefficients estimated via multinomial logit. Clustered standard errors at respondent level in parentheses. $\mathbb{1}_{\{X>\text{med}(X)\}}$ refers to respondents above the median on the respective attitude scale (AI Attitudes for columns 1-2, Science Attitudes for columns 3-4) The reference category for the synthesis method is the "Expert" produced synthesis status quo. The "Automatic AI" option refers to synthesis produced by a general-purpose model (*Generalist AI*) or a validated model (*Validated AI*). Total $N = 813$ respondents $\times$ 8 choice tasks $\times$ 2 alternatives. Asterisks indicate statistical significance: * $p < .05$, ** $p < .01$, *** $p < .001$.

Attitudes towards AI. Table 4 reports the results for heterogeneity by Attitudes towards AI, estimated separately by treatment arm. Estimates in Columns (1) and (2) reveal substantial heterogeneity. In the Control group (Column 1), policymakers with less positive attitudes towards AI (below the median) show a strong aversion towards a synthesis production approach fully automated by a general-purpose AI model ($\beta = -2.78, p < .001$) and do not display a preference for the Hybrid over the Expert approach. This indicates that the Hybrid approach is not always dominant over the status quo. Having more positive attitudes towards AI (i.e., being above the median) is associated with a reduced aversion towards the fully automated option and a strong preference for the Hybrid approach over the status quo. This suggests that for "AI optimists", general-purpose AI is already perceived as a useful productivity tool which, and when combined with expert oversight, can deliver higher utility than the expert-only status quo.

In the Treatment group (Column 2), policymakers with less positive attitudes show lower aversion to automation compared to their counterparts in the control group, with an estimate similar to that of respondents with more positive attitudes in the control group. Furthermore, in contrast to the control group, the coefficient on the Hybrid approach changes sign, indicating a slight preference over the status quo even among those more sceptical of AI. Policymakers with positive AI attitudes become more willing to substitute an expert-produced synthesis with one produced automatically by a validated AI model. These respondents are willing to pay a lower premium for the Hybrid approach over the status quo than the premium they would pay for the

status quo over the fully automated option. The results are similar, though less pronounced, when investigating heterogeneity by frequency of AI use.

Attitudes towards Science. Columns (3) and (4) in Table 4 report heterogeneity by Attitudes towards Science. As detailed in Section 2, this scale specifically measures normative expectations regarding the engagement of scientists in policymaking (e.g., "Scientists should work closely with policymakers") rather than trust in the scientific method per se. We observe fewer significant interaction terms compared to the AI attitudes heterogeneity analysis. In the Control group, the two interaction terms are imprecise. In contrast, in the Treatment group we identify a positive interaction effect for the Hybrid approach: respondents with more positive attitudes assign a higher premium to the Hybrid model relative to the status quo. The latter result may appear surprising, but the fact that the Hybrid model uses a scientifically validated AI model for evidence synthesis production and explicitly retains the expert (scientist) in the loop, may make it more appealing to those with more positive attitudes. It suggests that they do not view AI as a substitute for scientific authority but rather as a complementary tool, provided the tool is validated. Similarly, the absence of a negative interaction effect for the fully automated option, could be explained by the fact that, although scientists are absent from the final synthesis production, the validation of such models is still perceived as a process involving scientific expertise.

Role and Gender. Finally, Table 5 reports heterogeneity by gender and role. We observe no statistically significant heterogeneity along either dimension. Concerning gender, this finding is consistent with the descriptive statistics of our sample, which show no relevant differences in AI use or attitudes between male and female officials. Given that we have shown that attitudes predict behavioural intentions and choices, the absence of gender differences in baseline attitudes helps explain the absence of gender heterogeneity in stated preferences. This pattern presents a nuanced picture when compared to the broader literature on gender differences in AI adoption. Although several studies document gender gaps in general attitudes towards AI — with men reporting more positive views than women (Misra et al., 2024), greater willingness to support AI implementation in government agencies (Ahn and Chen, 2022), and greater resilience to pessimistic information about AI's labour market consequences (Huseynov, 2025) — other research suggests that these differences may not be universal or may not translate into differential behaviour in specific contexts. Giraldi et al. (2024), examining the Italian public sector, similarly found no significant gender differences in willingness to embrace new workplace technologies. Moreover, Misra et al. (2024) found that while gender predicted general positive attitudes towards AI, it was not a significant predictor of actual decisions between human assistance and AI tools in scenario-based choices—suggesting that even when attitudinal gaps exist, they may not translate into differential choices in specific implementation contexts. Similarly, the lack of differences between Political and Administrative roles goes against what is often found in the literature, which characterises administrators as technocratic consumers of evidence and politicians as more ideologically driven (Guridi et al., 2025). Our results suggest that preferences for the use of AI in evidence-synthesis production is

Table 5: Heterogeneity Analysis by Gender and Political Role

	Generalist	Validated	Generalist	Validated
Automatic AI	-2.212***	-1.279***	-2.234***	-1.351***
	(0.167)	(0.148)	(0.179)	(0.159)
Hybrid (AI + Expert)	0.225*	0.671***	0.239	0.731***
	(0.112)	(0.121)	(0.126)	(0.137)
AI $\times X$	-0.151	-0.041	-0.052	0.083
	(0.219)	(0.193)	(0.215)	(0.191)
Hybrid $\times X$	0.106	0.101	0.060	-0.040
	(0.179)	(0.180)	(0.169)	(0.176)
Cost (€000)	-0.277***	-0.257***	-0.277***	-0.252***
	(0.042)	(0.041)	(0.042)	(0.040)
Time (weeks)	-0.084***	-0.082***	-0.082***	-0.082***
	(0.019)	(0.019)	(0.019)	(0.019)
Log-Likelihood	-1556.9	-1641.3	-1565.2	-1651.8
$X =$	Female	Female	Pol. role	Pol. role

Notes: Coefficients estimated via multinomial logit. Clustered standard errors at respondent level in parentheses. X refers to being female for columns 1-2 and having a political role for columns 3-4. The reference category for the synthesis method is the "Expert" produced synthesis status quo. The "Automatic AI" option refers to synthesis produced by a general-purpose model (*Generalist AI*) or a validated model (*Validated AI*). Total $N = 813$ respondents $\times$ 8 choice tasks $\times$ 2 alternatives. Asterisks indicate statistical significance: * $p < .05$, ** $p < .01$, *** $p < .001$.

shared across the institutional spectrum.

Additional heterogeneity analyses are reported in the Online Appendix Tables A10 and A11. Table A10 examines heterogeneity by frequency of scientific evidence use, perceived usefulness of scientific evidence, and frequency of AI use, while Table A11 reports heterogeneity by age, education, and municipality size. These analyses complement the main heterogeneity results reported in Tables 5 and 6.

Mixed Logit Models. Mixed logit model results, reported in Table 6, reveal substantial heterogeneity in policymaker preferences, as evidenced by the highly significant standard deviations for all random parameters. For instance, in the Control group, while the mean preference for the Hybrid approach is positive, the large standard deviation ($sd = 2.46$) implies that a share of respondents in this group actually derive negative utility from the Hybrid model compared to the Expert status quo. This indicates that without specific validation, the preference for the hybrid production approach is driven by a subset of policymakers, while others remain sceptical. Treatment, however, fundamentally shifts this distribution. Exposure to the validated AI model not only increases the mean preference for the Hybrid approach but also tightens the distribution of preferences. Finally, allowing for random parameters on cost and delivery time reveals significant heterogeneity in sensitivity to efficiency constraints, suggesting that while some policymakers act under tighter budget and deadline constraints, others are relatively insensitive to these factors.

Table 6: Mixed Logit Estimates: Preference Heterogeneity

	Generalist AI	Validated AI
Mean Preferences		
Automatic AI	-5.729***	-3.091***
	(0.479)	(0.287)
Hybrid (AI + Expert)	0.624***	1.530***
	(0.147)	(0.163)
Cost	-0.687***	-0.621***
	(0.087)	(0.080)
Time (Mean)	-0.192***	-0.207***
	(0.035)	(0.035)
Unobserved Heterogeneity		
SD: Automatic AI	3.218***	2.378***
	(0.338)	(0.273)
SD: Hybrid	2.465***	1.990***
	(0.268)	(0.225)
SD: Cost	1.049***	1.040***
	(0.137)	(0.128)
SD: Time	0.350***	0.462***
	(0.054)	(0.053)
Log-Likelihood	-1290.0	-1384.4
McFadden R^2	0.451	0.358

Note: The table reports Mean (μ) and Standard Deviation (σ) parameters from Mixed Logit models with 1,000 Halton draws. All random parameters are assumed to follow a Normal distribution. The reference category is the "Expert" produced synthesis status quo. Total $N = 813$ respondents $\times$ 8 choice tasks $\times$ 2 alternatives. Asterisks indicate statistical significance: * $p < .05$, ** $p < .01$, *** $p < .001$.

4.3 Robustness checks

We performed several validity and robustness checks. In addition to the differential attrition and balance checks reported previously, we find that the randomisation produced an actual distribution of attribute levels close to the experimental design and that profile order (left vs right) has no effect on choices. Furthermore, we assess the sensitivity of the results to: (i) the exclusion of the few respondents ($N = 16$) who always chose the same synthesis production approach (Table A13, and (ii) the ordering of the AI and evidence-use survey sections (Table A14). Results for the former are consistent with our main findings, while for the latter, we find that being exposed to the AI section first appears to further increase the premium of the Hybrid approach. Finally, since the automatic AI option is somewhat dominated despite its lower price and delivery time, we investigate whether preferences for the hybrid approach relative to the expert approach are invariant if we exclude tasks including the automatic AI approach (Table A12). While the premium for the hybrid approach is attenuated and becomes insignificant in the Generalist AI control group (partly due to the reduction in sample size), none of the main results is affected.

5 Conclusion

This paper investigates policymakers' preferences for the use of generative artificial intelligence in the production of scientific evidence syntheses for policy design. Using a discrete choice experiment with more than 800 Italian policymakers, we quantify how they trade off different evidence synthesis production approaches, delivery time, and cost. Furthermore, in a between-subject design, we investigate how these preferences respond to whether the AI model used is described as a general-

purpose model or as specifically validated for evidence synthesis. The analysis provides one of the first empirical assessments of demand-side preferences for AI-assisted knowledge production in public policy.

Three main conclusions emerge. First, policymakers show a strong baseline aversion to fully automated evidence synthesis, even when automation delivers substantial efficiency gains, possibly reflecting concerns about reliability, hallucinations, and the loss of expert judgment. Task-specific validation of AI systems significantly increases acceptance, particularly by reducing aversion among more sceptical policymakers, but on average policymakers still prefer expert-produced evidence when cost and time are held constant. This persistent resistance suggests that concerns about AI in high-stakes policy contexts cannot be fully addressed through technical validation alone. Rather, they may reflect deeper considerations about epistemic authority, accountability, and the appropriate role of algorithmic systems in shaping policy decisions.

Second, irrespective of whether the AI model is general purpose or validated, a hybrid approach, combining AI-generated drafts with expert revision, is systematically preferred over both expert-only and AI-only approaches, conditional on cost and delivery time. Task-specific validation further strengthens support. These results indicate that policymakers value AI primarily as a productivity-enhancing complement to human expertise, rather than as a replacement for it. The preference for the hybrid approach may reflect perceived value in human-AI complementarity: policymakers may believe that AI can expand the scope and comprehensiveness of evidence review while expert oversight ensures quality and appropriate interpretation.

Third, we find strong heterogeneity in preference. Preferences vary systematically with attitudes towards AI, where sceptics require validation before accepting even hybrid models, while those with positive AI attitudes show willingness to use general-purpose AI with expert oversight. Interestingly, we find no significant heterogeneity by gender or institutional role.

Our findings indicate that acceptance of AI integration in evidence synthesis for the public administration requires both task-specific validation and expert oversight, pointing towards a complementary hybrid human-AI workflow rather than full automation. The results have direct implications for AI governance frameworks, evidence synthesis production, and the broader adoption of AI tools in knowledge-intensive public sector tasks, highlighting the central role of human oversight and task-specific validation in developing trust and legitimacy.

The study has several limitations. First, a limitation of the study in terms of understanding of adoption decisions is that we elicit the preferences of policymakers, but actual adoption decisions may differ once institutional or legal constraints come into play. While the survey was 100% anonymous and we used a neutral language, we cannot discard potential social desirability effects, though the expected direction of such effects is not obvious, as the topic under investigation is not particularly normatively charged. Second, the sample, while large and institutionally diverse, cannot be assumed to be fully representative of the population of Italian policymakers. Our sampling strategy was designed to maximise institutional coverage, and the analytic sam-

ple closely matches available administrative benchmarks on the dimensions that can be compared, such as gender and age distributions for local and regional politicians and administrative managers. However, comparable population-level information is not available for all relevant characteristics, roles, territorial levels, or policy domains. We therefore cannot guarantee representativeness along dimensions that are unobserved or not externally benchmarkable. Third, to avoid cognitive fatigue, our design abstracts from other relevant attributes, important to policymakers when deciding the usefulness of an evidence synthesis in policy design.

Despite these limitations, the findings have important policy implications. For researchers, governments, and private actors seeking to integrate AI into evidence-based policymaking, the results suggest that fully automated systems are unlikely to gain legitimacy in high-stakes contexts. Instead, investments should prioritise hybrid human–AI models, coupled with transparent validation and task-specific evaluation of AI models. More broadly, the results support regulatory approaches that emphasise validation-for-use and human oversight as central safeguards for AI deployment in public decision-making.

Declaration on the Use of AI Tools

During the preparation of this manuscript, the authors used AI-based tools (e.g., large language models such as Claude and ChatGPT, and grammar-checking tools such as Grammarly) solely to improve clarity, structure, or readability. All content has been independently reviewed, verified, and edited by the authors, who take full responsibility for the accuracy, originality, and integrity of the manuscript.

References

Adel, A. and Alani, N. (2025). Can generative AI reliably synthesise literature? Exploring hallucination issues in ChatGPT. *AI & Society*, 40:6799–6812.

Ahn, M. J. and Chen, Y.-C. (2022). Digital transformation toward AI-augmented public administration: The perception of government employees and the willingness to use AI in government. *Government Information Quarterly*, 39(2):101664.

Alon-Barkat, S. and Busuioc, M. (2023). Human–AI interactions in public sector decision making: "Automation bias" and "Selective adherence" to algorithmic advice. *Journal of Public Administration Research and Theory*, 33(1):153–169.

Angrist, J. D. and Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24(2):3–30.

Aoki, N. (2021). The importance of the assurance that "humans are still in the decision loop" for public trust in artificial intelligence. *Computers in Human Behavior*, 114:106572.

Banuri, S., Dercon, S., and Gauri, V. (2019). Biased policy professionals. *The World Bank Economic Review*, 33(2):310–327.

Bommasani, R. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Bortolus, A., Hewitt, C. L., Mitova, V., et al. (2025). Knowledge brokers at the science–policy interface: Insights from biosecurity and environmental management. *Ambio*, 54:2079–2091.

Briscese, G. and List, J. A. (2024). Toward an understanding of the political economy of using field experiments in policymaking. NBER Working Paper No. 33239.

Cologna, V., Mede, N. G., Berger, S., Besley, J., Brick, C., Joubert, M., Maibach, E. W., Mihelj, S., Oreskes, N., Schäfer, M. S., et al. (2025). Trust in scientists and their role in society across 68 countries. *Nature Human Behaviour*, 9(4):713–730.

Cong-Lem, N. (2025). Rethinking evidence-informed policy and practice in the age of generative artificial intelligence. *London Review of Education*, 23(1):16.

Currie, J., Kleven, H., and Zwiers, E. (2020). Technology and big data are changing economics: Mining text to track methods. In *AEA Papers and Proceedings*, volume 110, pages 42–48. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

DellaVigna, S., Kim, W., and Linos, E. (2024). Bottlenecks for evidence adoption. *Journal of Political Economy*, 132(8):2748–2789.

Dhar, R. and Simonson, I. (2003). The effect of forced choice on choice. *Journal of Marketing Research*, 40(2):146–160.

European Commission (2024). Artificial intelligence in the European Commission (AI@EC) a strategic vision to foster the development and use of lawful, safe and trustworthy Artificial Intelligence systems in the European Commission. *European Commission*, (349).

European Union (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence.

Flemyng, E., Noel-Storr, A., Macura, B., Gartlehner, G., Thomas, J., Meerpohl, J. J., Jordan, Z., Minx, J., Eisele-Metzger, A., Hamel, C., et al. (2025). Position statement on artificial intelligence (AI) use in evidence synthesis across Cochrane, the Campbell Collaboration, JBI and the Collaboration for Environmental Evidence 2025. *Cochrane Database of Systematic Reviews*, (10).

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., et al. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4):689–707.

Gieselmann, M., Erdsiek, D., Rost, V., and Sassenberg, K. (2025). Do managers accept artificial intelligence? Insights into the role of business area and AI functionality. *Journal of Economic Psychology*, 108:102804.

Giraldi, L., Rossi, L., and Rudawska, E. (2024). Evaluating public sector employee perceptions towards artificial intelligence and generative artificial intelligence integration. *Journal of Information Science*, page 01655515241293775.

Gnambs, T., Stein, J.-P., Appel, M., Griese, F., and Zinn, S. (2025). An economical measure of attitudes towards artificial intelligence in work, healthcare, and education (ATTARI-WHE). *Computers in Human Behavior: Artificial Humans*, 3:100106.

Goldsmith-Pinkham, P. (2024). Tracking the credibility revolution across fields. *arXiv preprint arXiv:2405.20604*.

Guridi, J. A., Cheyre, C., and Yang, Q. (2025). Thoughtful adoption of NLP for civic participation: Understanding differences among policymakers. *Proceedings of the ACM on Human-Computer Interaction*, 9(2).

Hawkins, D., Beesley, C., Nielson, D., Lyne, M., Morgenstern, S., Garcia, G., Walker, L., and Long, K. (2024). Closing the gap between evidence and policy in Latin America: What works? Preprint, Research Square.

Hill, H. C. and Briggs, D. C. (2020). Education leaders' knowledge of causal research design: A measurement challenge. Technical Report EdWorkingPaper 20-298, Annenberg Institute at Brown University.

Hillo, J., Vento, I., and Erkkilä, T. (2026). Algorithmic governance: Experimental evidence on citizens' and public administrators' legitimacy perceptions of automated decision-making. *Public Administration*, 104:536–551.

Hjort, J., Moreira, D., Rao, G., and Santini, J. F. (2021). How research affects policy: Experimental evidence from 2,150 Brazilian municipalities. *American Economic Review*, 111(5):1442–1480.

Huseynov, S. (2025). ChatGPT and the labor market: Unraveling the effect of AI discussions on students' earning expectations. *Journal of Economic Psychology*, 108:102803.

Johnston, R. J., Boyle, K. J., Adamowicz, W., Bennett, J., Brouwer, R., Cameron, T. A., Hanemann, W. M., Hanley, N., Ryan, M., Scarpa, R., et al. (2017). Contemporary guidance for stated preference studies. *Journal of the Association of Environmental and Resource Economists*, 4(2):319–405.

Levy, K., Chasalow, K. E., and Riley, S. (2021). Algorithms and decision-making in the public sector. *Annual Review of Law and Social Science*, 17:309–334.

Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103.

Lupia, A. and McCubbins, M. D. (1998). *The Democratic Dilemma: Can Citizens Learn What They Need to Know?* Cambridge University Press.

McFadden, D. (2001). Economic choices. *American Economic Review*, 91(3):351–378.

Misra, S., Katz, B., Roberts, P., Carney, M., and Valdivia, I. (2024). Toward a person-environment fit framework for artificial intelligence implementation in the public sector. *Government Information Quarterly*, 41(3):101962.

Nakajima, N. (2021). Evidence-based decisions and education policymakers. Technical report, Working paper.

National Institute of Standards and Technology (2023). Artificial intelligence risk management framework (AI RMF 1.0). Technical report, U.S. Department of Commerce.

Newman, J. and Mintrom, M. (2023). Mapping the discourse on evidence-based policy, artificial intelligence, and the ethical practice of policy analysis. *Journal of European Public Policy*, 30(9):1839–1859.

OECD (2020). *Building Capacity for Evidence-Informed Policy-Making: Lessons from Country Experiences*. OECD Publishing, Paris.

OECD (2023). Advancing accountability in AI: Governing and managing risks throughout the lifecycle for trustworthy AI. *OECD Digital Economy Papers*, (349).

Oliver, K., Innvar, S., Lorenc, T., Woodman, J., and Thomas, J. (2014). A systematic review of barriers to and facilitators of the use of evidence by policymakers. *BMC Health Services Research*, 14:Article 2.

Pereira, M. M., Baekgaard, M., Giger, N., and Sheffer, L. (2024). Politicians and evidence. Technical report, OSF Preprints.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 33–44.

Rao, M. (2024). Program evaluations and policy spending. Technical report, Working paper.

Speckemeier, C., Krabbe, L., Schwenke, S., Wasem, J., Buchberger, B., and Neusser, S. (2021). Discrete choice experiment to determine preferences of decision-makers in healthcare for different formats of rapid reviews. *Systematic Reviews*, 10:121.

Tangi, L., Rodriguez Müller, A. P., and Combetto, M. (2026). A silent partner: The shadow presence of generative artificial intelligence in public administrations. In *Electronic Participation*, volume 15978 of *Lecture Notes in Computer Science*. Springer, Cham.

Toma, M. and Bell, E. (2024). Understanding and increasing policymakers' sensitivity to program impact. *Journal of Public Economics*, 234:105096.

Vivalt, E. and Coville, A. (2023). How do policymakers update their beliefs? *Journal of Development Economics*, 165:103121.

Vivalt, E., Coville, A., and KC, S. (2025). Local knowledge, formal evidence, and policy decisions. *Journal of Development Economics*, 174:103425.

Voutyrakou, D. A. and Skordoulis, C. (2025). Algorithmic governance: Gender bias in AI-generated policymaking? *Human-Centric Intelligent Systems*, 5:385–417.

Wirtz, B. W., Weyerer, J. C., and Geyer, C. (2019). Artificial intelligence and the public sector—applications and challenges. *International Journal of Public Administration*, 42(7):596–615.

Xu, C., Cheng, Y. D., Wang, S., Merrick, W., and Carter, P. (2024). Evaluating use of evidence in U.S. state governments: A conjoint analysis. *Public Administration Review*, pages 1–19.

Online Appendix: GenAI: Substitute or Complement in Evidence Synthesis. A Discrete Choice Experiment with Italian Policymakers

Martina Bazzoli, Sergiu Burlacu, Dominique Cappelletti, Sonia Marzadro, Iunio Quarto Russo, Alessio Tomelleri, and Mirco Tonin

Supplementary Analyses

Table A1: Survey ($N = 413$) and administrative statistics for local and regional politicians

	Sample % (1)	Admin. % (2)
Gender		
Female	39.0	34.2
Male	60.3	65.7
Age group		
18 to 35	16.2	16.4
36 to 55	45.3	47.8
56 to 65	22.3	21.7
Over 65	16.0	13.9
Education		
High school or less	37.5	46.8
University degree	45.8	29.3
Post-graduate degree	16.0	0.97
Not provided	0.7	22.8

Note: Column (1) reports survey data statistics while Column (2) reports official statistics, for local and regional politicians. *Source for admin. data:* Registry of Local and Regional Elected Officials (it *"Anagrafe degli amministratori locali e regionali"*), Department for Internal and Territorial Affairs, Ministry of the Interior. Percentages may not sum to exactly 100% due to rounding or the exclusion of residual categories.

Table A2: Survey ($N = 277$) and administrative statistics for administrative managers

	Sample % (1)	Admin. % (2)
Gender		
Female	37.5	40.1
Male	62.1	59.9
Age group*		
30 to 55	42.2	40.4
Over 56	56.3	59.6
Education		
High school or less	1.4	
University degree	58.1	
Post-graduate degree	40.4	

Note: Column (1) reports survey data statistics (year 2025) while Column (2) reports official administrative statistics for municipal managers (year 2023). The administrative statistics are only for managers with permanent contract, while the survey data may include also fixed term or flexible contracts. *Source for admin. data:* IFEL-Ufficio Studi e Statistiche Territoriali processing of Ministry of Economy and Finance data, 2025. Percentages may not sum to exactly 100% due to rounding or the exclusion of residual categories. * The administrative age-group statistics are based on slightly different intervals: "30-54" and "55 and above". The administratie data covers only municipal managers (ït *"dirigenti comunali"*, while the survey data may include other managerial roles (it *"direttori, quadri"*, and it consider also Province and Regional level managers. Results are similar if restricting the survey data to the smaller sample of municipal mangers ($N = 116$).

Table A3: Completing and drop-out rates by experimental stage

	N	%
Completed	813	0.91
Dropped at intro	38	0.04
Dropped during tasks	47	0.05

Notes: Dropped at intro indicates respondents dropping out at the instructional page of the experiment. Dropped during tasks indicates respondents dropping out during the experimental tasks.

Table A4: Attrition by between-subject treatment arm

	Attrition
Validated AI	0.013
	(0.020)
Num.Obs.	898
Generalist AI mean	0.088

Notes: Coefficients estimated via OLS. Dependent variable is an indicator for drop-out at any stage of experimental task. Reference category is the Generalist AI arm. Heteroscedasticity robust standard errors in parentheses. Asterisks indicate statistical significance: * $p < .05$, ** $p < .01$, *** $p < .001$.

Table A5: Attrition by respondent characteristics and treatment status

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Validated AI	0.010	0.006	0.007	-0.001	0.013	0.014	0.014
	(0.026)	(0.028)	(0.027)	(0.039)	(0.026)	(0.020)	(0.019)
X	0.018	0.000	0.014	-0.047	0.015	-0.010	0.010
	(0.028)	(0.027)	(0.028)	(0.031)	(0.027)	(0.013)	(0.014)
$X\times$ Validated AI	0.007	0.015	0.011	0.018	0.002	-0.007	-0.056**
	(0.040)	(0.039)	(0.039)	(0.045)	(0.039)	(0.020)	(0.019)
Num.Obs.	893	898	894	890	897	898	898
$X =$	Female	Pol. role	Age ≥ 56	Uni. +	Mun. < 20k	Science att.	AI att.

Notes: Coefficients estimated via OLS. The reference category is the Generalist AI arm. AI attitudes and Science attitudes are standardised with mean 0 and standard deviation 1. Uni. + indicates having a university degree or higher. Mun. < 20k indicates operating in a municipality with population under 20,000. Pol. role indicates having a political role (ref. administrative). Heteroscedasticity robust standard errors in parentheses. Asterisks indicate statistical significance: * $p < .05$, ** $p < .01$, *** $p < .001$. Omnibus F-test of joint significance of all interaction terms included in a joint model: F = 0.01, p = .906.

Table A6: Balance of demographics by between-subject treatment arm

	N	Generalist AI	Validated AI	SMD	p-value
Role within public administration (%)					
Political role	813	0.51	0.50	-0.02	0.82
Administrative role	813	0.49	0.50	0.02	0.82
Territorial jurisdiction (%)					
Municipal (fewer than 2,000 inhabitants)	813	0.12	0.12	-0.03	0.68
Municipal (2,000 to 20,000 inhabitants)	813	0.34	0.31	-0.06	0.43
Municipal (over 20,000 inhabitants)	813	0.28	0.29	0.01	0.94
Province / Metropolitan City	813	0.09	0.08	-0.05	0.52
Regional	813	0.15	0.20	0.12	0.08
Geographic area (%)					
North-East	807	0.46	0.47	0.01	0.87
North-West	807	0.31	0.28	-0.05	0.48
Center	807	0.14	0.16	0.05	0.49
South	807	0.05	0.04	-0.03	0.68
Islands	807	0.04	0.05	0.03	0.67
Gender (%)					
Female	813	0.38	0.43	0.09	0.19
Male	813	0.61	0.56	-0.10	0.17
Age group (%)					
18 to 35	813	0.10	0.11	0.01	0.94
36 to 55	813	0.45	0.44	-0.01	0.86
56 to 65	813	0.33	0.38	0.10	0.16
Over 65	813	0.11	0.08	-0.11	0.11
Education (%)					
High school or less	813	0.24	0.23	-0.03	0.64
University degree	813	0.53	0.52	-0.02	0.74
Post-graduate degree	813	0.23	0.25	0.04	0.55
Attitudes towards science $[0, 1]$	813	0.72	0.73	0.05	0.44
Attitudes towards AI $[0, 1]$	813	0.58	0.59	0.08	0.24

Notes: The table reports baseline means for the control group *Generalist AI* and treatment group (*Validated AI*). SMD indicates standardised mean differences. The last column reports p-values from regression models of the baseline characteristics on treatment status, using heteroscedasticity robust standard errors. Residual categories (e.g., "Other" or "Prefer not to respond") are not reported in the table, which may cause some sections not to sum to exactly 100%. Attitudes towards AI and science are normalised on a 0-1 scale, where 0 and 1 indicate the theoretical minimum and maximum, respectively. All composite indices are codes such that higher values indicate more positive attitudes.

Table A7: Willingness-to-pay (€) estimates

Attribute	WTP	SE	95% CI
Full sample			
Automatic AI	-6758	(693)	[-8116, -5399]
Hybrid (AI + Expert)	1874	(357)	[1174, 2574]
Time (per week)	-311	(60)	[-428, -194]
Generalist AI			
Automatic AI	-8155	(1153)	[-10415, -5895]
Hybrid (AI + Expert)	970	(389)	[207, 1734]
Time (per week)	-297	(81)	[-456, -137]
Validated AI			
Automatic AI	-5186	(787)	[-6728, -3644]
Hybrid (AI + Expert)	2816	(639)	[1564, 4068]
Time (per week)	-325	(88)	[-497, -153]

Notes: WTP computed by dividing with the cost coefficient estimate. Delta-method SEs reported in Column SE in parentheses, 95% CI in brackets. The reference category for the synthesis method is the "Expert" produced synthesis status quo. The "Automatic AI" option refers to synthesis produced by a general-purpose model (*Generalist AI*) or a validated model (*Validated AI*).

Table A8: Preferences for Evidence Synthesis - Treatment-interaction model

	(1)
Automatic AI	-2.242***
	(0.124)
Hybrid (AI + Expert)	0.277**
	(0.089)
Automatic AI × Validated AI	0.917***
	(0.144)
Hybrid × Validated AI	0.427***
	(0.122)
Cost (€000)	-0.264***
	(0.029)
Time (weeks)	-0.082***
	(0.013)

Notes: Coefficients estimated via multinomial logit. Clustered standard errors at the respondent level in parentheses. Coefficients from multinomial logit models. The reference category for the synthesis method is the "Expert" produced synthesis status quo. The "Automatic AI" option refers to synthesis produced by a general-purpose model (*Generalist AI*) or a validated model (*Validated AI*). Total $N = 813$ respondents × 8 choice tasks × 2 alternatives. Asterisks indicate statistical significance: * p < .05, ** p < .01, *** p < .001.

Table A9: Linear tests of differences between coefficients

Test	Null hypothesis	Estimate	Cluster SE	Wald $\chi^2(1)$	p-value
1. Hybrid vs. AI-only, Control	$\beta_2 - \beta_1 = 0$	2.519	0.107	552.6	< 0.001
2. Hybrid vs. AI-only, Treatment	$(\beta_2 + \delta_2) - (\beta_1 + \delta_1) = 0$	2.029	0.090	513.2	< 0.001
3. Effect of validation: AI-only vs. Hybrid	$\delta_1 - \delta_2 = 0$	0.490	0.133	13.6	< 0.001

Notes: Linear combinations of the coefficients of the pooled model in Equation (3) (Table A8), where β_1 and β_2 refer to the AI-only and Hybrid options relative to expert production in the Control group, and δ_1 and δ_2 to their interactions with the Validated AI condition. Standard errors are clustered at the respondent level; the test statistic is a Wald χ^2 with one degree of freedom. By construction, Test 2 equals Test 1 minus Test 3.

Table A10: Heterogeneity Analysis by frequency of AI and scientific evidence use, and perceived usefulness of evidence

	Generalist	Validated	Generalist	Validated	Generalist	Validated
Automatic AI	-2.255***	-1.089***	-1.659***	-0.743*	-2.453***	-1.611***
	(0.216)	(0.186)	(0.339)	(0.320)	(0.199)	(0.157)
Hybrid (AI + Expert)	0.137	0.438**	0.004	0.617	-0.043	0.524***
	(0.150)	(0.164)	(0.282)	(0.328)	(0.130)	(0.126)
AI $\times X$	-0.014	-0.305	-0.683*	-0.619	0.342	0.611**
	(0.231)	(0.211)	(0.344)	(0.335)	(0.219)	(0.192)
Hybrid $\times X$	0.187	0.380*	0.291	0.100	0.612***	0.393*
	(0.179)	(0.193)	(0.294)	(0.340)	(0.169)	(0.176)
Cost (€000)	-0.278***	-0.254***	-0.281***	-0.250***	-0.279***	-0.258***
	(0.042)	(0.041)	(0.042)	(0.041)	(0.042)	(0.041)
Time (weeks)	-0.082***	-0.082***	-0.083***	-0.084***	-0.083***	-0.080***
	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)
Log-Likelihood	-1564.2	-1640.9	-1555.9	-1646.3	-1551.2	-1637.7
$X =$	Science use	Science use	Science useful	Science useful	AI use	AI use

Notes: Coefficients estimated via multinomial logit. Clustered standard errors at respondent level in parentheses. The reference category for the synthesis method is the "Expert" produced synthesis status quo. The "Automatic AI" option refers to synthesis produced by a general-purpose model (*Generalist AI*) or a validated model (*Validated AI*). Total $N = 813$ respondents $\times$ 8 choice tasks $\times$ 2 alternatives. Asterisks indicate statistical significance: * p $< .05$, ** p $< .01$, *** p $< .001$

Table A11: Heterogeneity Analysis by Age, Education and Municipality size

	Generalist	Validated	Generalist	Validated	Generalist	Validated
Automatic AI	-2.272***	-1.276***	-1.861***	-1.066***	-2.504***	-1.535***
	(0.159)	(0.148)	(0.216)	(0.200)	(0.177)	(0.155)
Hybrid (AI + Expert)	0.258*	0.646***	0.217	0.588***	0.405**	0.758***
	(0.115)	(0.122)	(0.176)	(0.163)	(0.124)	(0.130)
AI $\times X$	-0.018	-0.077	-0.577*	-0.333	0.440*	0.482*
	(0.220)	(0.193)	(0.232)	(0.219)	(0.216)	(0.193)
Hybrid $\times X$	0.008	0.143	0.063	0.162	-0.287	-0.094
	(0.173)	(0.177)	(0.195)	(0.188)	(0.170)	(0.177)
Cost (€000)	-0.278***	-0.252***	-0.279***	-0.255***	-0.282***	-0.255***
	(0.042)	(0.040)	(0.042)	(0.041)	(0.042)	(0.041)
Time (weeks)	-0.085***	-0.083***	-0.085***	-0.084***	-0.083***	-0.085***
	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)
Log-Likelihood	-1546.8	-1650.8	-1555.6	-1632.6	-1552.6	-1635.6
$X =$	Age $\geq=$ 56	Age $\geq=$ 56	Uni. +	Uni. +	Municipality $<$ 20k	Municipality $<$ 20k

Notes: Coefficients estimated via multinomial logit. Clustered standard errors at respondent level in parentheses. The reference category for the synthesis method is the "Expert" produced synthesis status quo. The "Automatic AI" option refers to synthesis produced by a general-purpose model (*Generalist AI*) or a validated model (*Validated AI*). Total $N = 813$ respondents $\times$ 8 choice tasks $\times$ 2 alternatives. Asterisks indicate statistical significance: * p $<$ .05, ** p $<$ .01, *** p $<$ .001.

Table A12: Robustness check: multinomial logit estimates excluding the automatic AI alternative

	Full Sample	Generalist AI	Validated AI
Hybrid (AI + Expert)	0.316***	0.183	0.467***
	(0.073)	(0.103)	(0.105)
Cost (€000)	-0.231***	-0.234***	-0.225***
	(0.038)	(0.053)	(0.054)
Time (weeks)	-0.080***	-0.069**	-0.089***
	(0.017)	(0.023)	(0.025)
Log-Likelihood	-1364.0	-717.6	-640.0

Notes: Coefficients estimated via multinomial logit. Clustered standard errors at respondent level in parentheses. The reference category for the synthesis method is the "Expert" produced synthesis status quo. The sample excludes choice tasks with the Automatic AI option (2181 of 6504 choice tasks). Asterisks indicate statistical significance: * p $<$ .05, ** p $<$ .01, *** p $<$ .001.

Table A13: Robustness check: excluding respondents always choosing the same production type option

	Full Sample	Generalist AI	Validated AI
Automatic AI	-1.760***	-2.266***	-1.306***
	(0.092)	(0.139)	(0.121)
Hybrid (AI + Expert)	0.455***	0.244**	0.668***
	(0.067)	(0.093)	(0.097)
Cost (€000)	-0.262***	-0.282***	-0.251***
	(0.029)	(0.042)	(0.041)
Time (weeks)	-0.082***	-0.084***	-0.083***
	(0.013)	(0.019)	(0.019)
Log-Likelihood	-3240.5	-1548.2	-1636.4

Notes: Coefficients estimated via multinomial logit. Clustered standard errors at respondent level in parentheses. The reference category for the synthesis method is the "Expert" produced synthesis status quo. The "Automatic AI" option refers to synthesis produced by a general-purpose model (*Generalist AI*) or a validated model (*Validated AI*). The sample excludes respondents who always choose the same production type option (N = 16 out of 813). Asterisks indicate statistical significance: * p < .05, ** p < .01, *** p < .001.

Table A14: Robustness check: order effects of AI vs evidence use survey section

	Full Sample	Generalist AI	Validated AI
Automatic AI	-1.785***	-2.144***	-1.404***
	(0.113)	(0.163)	(0.154)
Hybrid (AI + Expert)	0.388***	0.295*	0.497***
	(0.088)	(0.122)	(0.128)
Automatic AI × AI section first	0.055	-0.244	0.203
	(0.143)	(0.217)	(0.193)
Hybrid × AI section first	0.205	-0.052	0.430*
	(0.121)	(0.170)	(0.176)
Cost (€000)	-0.260***	-0.277***	-0.248***
	(0.029)	(0.042)	(0.041)
Time (weeks)	-0.081***	-0.081***	-0.081***
	(0.013)	(0.019)	(0.019)
Log-Likelihood	-3271.4	-1563.8	-1645.8

Notes: Coefficients estimated via multinomial logit. Clustered standard errors at respondent level in parentheses. The reference category for the synthesis method is the "Expert" produced synthesis status quo. The "Automatic AI" option refers to synthesis produced by a general-purpose model (*Generalist AI*) or a validated model (*Validated AI*). AI section first indicates if the survey section on AI use and attitudes was randomly presented before the section evidence use and attitudes towards science. Total $N = 813$ respondents × 8 choice tasks × 2 alternatives. A joint Wald test of the null that both interactions with the AI-section-first indicator are zero yields p = .201 for the full sample, p = .515 for the Generalist AI condition, and p = .047 for the Validated AI condition. Asterisks indicate statistical significance: * p < .05, ** p < .01, *** p < .001.

Table A15: Mapping between the pre-analysis plan (AEARCTR-0016366) and the analyses in the paper

PAP item	Status in PAP	Where in the paper	Classification
Choice probability by production method	Primary	Figure 1; Table 3	Pre-registered (primary)
Effect of validated vs. generalist AI on AI preference	Primary	Table 3 (cols. 2–3); Tables A8, A9	Pre-registered (primary)
Willingness to pay by production method	Primary	Table A7	Pre-registered (primary)
Marginal rates of substitution (cost vs. time)	Primary	Table A7 (WTP per week)	Pre-registered (primary)
Attribute importance weights	Primary	Table A22	Pre-registered (primary)
Heterogeneity: attitudes towards AI and science	Secondary	Table 4	Pre-registered (F1)
Heterogeneity: political vs. administrative role	Secondary	Table 5	Pre-registered (F2)
Heterogeneity: government level	Secondary	Table A11 (municipality size)	Pre-registered (F2)
Heterogeneity: demographics (gender, age, education)	Secondary	Tables 5, A11	Pre-registered (F2)
Heterogeneity: geographic area	Secondary	Tables A23–A24	Pre-registered (F2)
Heterogeneity: policy areas	Secondary	Tables A17–A18 (MHT table)	Pre-registered (F2b)
Correlation with AI use and evidence use	Secondary	Table A10	Pre-registered (F2)
Preference classes (latent class)	Secondary	Table A25	Pre-registered
Mixed logit	Not in PAP	Table 6	Exploratory (unobserved heterogeneity)
Attrition, balance, design checks	Not in PAP	Tables A3–A6	Standard validity checks
Hybrid premium without AI / by pair type	Not in PAP	Table A12; Table A26	Robustness (not pre-registered)
Excluding respondents always choosing the same method	Not in PAP	Table A13	Robustness (not pre-registered)
Order of AI and evidence survey sections	Design feature; analysis not in PAP	Table A14	Robustness (not pre-registered)
Heterogeneity by policy tasks / commissioning proxies	Not in PAP	Tables A20–A21	Exploratory (F3)
Individual conditional means (mixed logit)	Not in PAP	Table A27	Exploratory (unobserved heterogeneity)
Willingness to pay relative to design support	Not in PAP	Table A28	Robustness (not pre-registered)
Predicted choice probabilities within the design	Not in PAP	Table A29	Exploratory (magnitudes)
Sensitivity to selection (more AI-sceptical sub-samples)	Not in PAP	Table A30	Robustness (not pre-registered)

Notes: F1–F3 refer to the families used for the multiple-hypothesis correction (Tables A17 and A18): F1, main heterogeneity dimensions (attitudes towards AI and science); F2, other pre-registered dimensions; F2b, policy areas; F3, policy tasks. Deviations from the PAP: (i) the achieved sample (N = 813) is below the target of 1,000–1,300; (ii) continuous moderators are split at the median and categorical ones grouped into two balanced categories, which the PAP did not specify.

Table A16: Constructs, measurement and empirical role

Construct	How it enters the study	Measure	Empirical use
Preferences over production method, cost and time	Elicited (stated preferences)	DCE: 8 binary choice tasks	Outcome (MNL, mixed logit, latent class)
Response to task-specific validation information	Experimentally manipulated	Between-subject description of the AI model	Treatment (interaction with AI-based options)
General attitudes towards AI at work	Measured	ATTARI-WORK, 3 items (Gnambs et al. 2025)	Moderator; latent-class membership
Normative attitudes towards science in policy	Measured	3 items from Cologna et al. (2025)	Moderator
GenAI use; use and perceived usefulness of scientific evidence	Measured (self-reported)	Single frequency / usefulness items	Moderators
Role, level of government, tasks performed, demographics	Measured (self-reported)	Background questionnaire	Moderators; descriptives
Algorithm appreciation, algorithm aversion (including task-dependent aversion)	Not measured	-	Interpretation only
Factors related to the quality of the output: perceived unreliability, opacity (technoskepticism), mind perception, uniqueness neglect	Not measured	-	Interpretation only
Factors related to the source of the synthesis: human favouritism, status quo bias, accountability (legitimacy, epistemic authority; blame-shifting vs. abdication)	Not measured	-	Interpretation only
Role of AI relative to experts: reassurance vs. complementarity	Not measured	-	Interpretation only

Table A17: Omnibus tests of heterogeneity with multiple-testing adjustment

Family	Dimension	Wald chi2(4)	p	Holm p	BH p
F1 Main heterogeneity	AI attitudes	66.484	0.000	0.000	0.000
F1 Main heterogeneity	Science attitudes	13.786	0.008	0.008	0.008
F2 Secondary (pre-registered)	Pol. role	0.856	0.931	1.000	0.931
F2 Secondary (pre-registered)	Municipality $<$ 20k	25.725	0.000	0.000	0.000
F2 Secondary (pre-registered)	North	5.509	0.239	0.956	0.358
F2 Secondary (pre-registered)	Female	2.229	0.694	1.000	0.871
F2 Secondary (pre-registered)	Age $>=$ 56	1.789	0.775	1.000	0.871
F2 Secondary (pre-registered)	Uni. +	14.563	0.006	0.028	0.010
F2 Secondary (pre-registered)	AI use	23.523	0.000	0.001	0.000
F2 Secondary (pre-registered)	Science use	16.059	0.003	0.020	0.007
F2 Secondary (pre-registered)	Science useful	16.175	0.003	0.020	0.007
F2b Policy areas (pre-registered)	Economy and labour	5.239	0.264	1.000	0.422
F2b Policy areas (pre-registered)	Education and research	2.253	0.689	1.000	0.781
F2b Policy areas (pre-registered)	Health	8.120	0.087	0.698	0.348
F2b Policy areas (pre-registered)	Social policies	3.552	0.470	1.000	0.627
F2b Policy areas (pre-registered)	Environment and sustainability	6.357	0.174	0.931	0.348
F2b Policy areas (pre-registered)	Security, justice, foreign policy, defence	6.664	0.155	0.931	0.348
F2b Policy areas (pre-registered)	Infrastructure and transport	7.054	0.133	0.931	0.348
F2b Policy areas (pre-registered)	Culture, tourism and sport	1.753	0.781	1.000	0.781
F3 Exploratory (tasks)	Decision-maker	1.029	0.905	1.000	0.905
F3 Exploratory (tasks)	Analyst	5.154	0.272	0.816	0.816
F3 Exploratory (tasks)	Tasks $>$ median	1.823	0.768	1.000	0.905

Notes: Wald test that the four interaction terms (Automatic AI $\times X$ and Hybrid $\times X$ in the Generalist and Validated AI arms) are jointly zero. The two arms are independent samples, so the joint covariance matrix is block-diagonal in the arm-specific cluster-robust covariance matrices. Holm (family-wise error rate) and Benjamini-Hochberg (false discovery rate) adjustments are applied within family.

Table A18: Heterogeneity interaction coefficients with multiple-testing adjustment

Family	Dimension	Arm	Term	Estimate	SE	p	Holm p	Sharpened q
F1	AI attitudes	Generalist	AI x X	0.891	0.219	0.000	0.000	0.001
F1	AI attitudes	Generalist	Hybrid x X	0.842	0.168	0.000	0.000	0.001
F1	AI attitudes	Validated	AI x X	1.127	0.197	0.000	0.000	0.001
F1	AI attitudes	Validated	Hybrid x X	0.862	0.177	0.000	0.000	0.001
F1	Science attitudes	Generalist	AI x X	-0.396	0.216	0.066	0.198	0.035
F1	Science attitudes	Generalist	Hybrid x X	0.106	0.171	0.536	1.000	0.155
F1	Science attitudes	Validated	AI x X	0.126	0.192	0.513	1.000	0.155
F1	Science attitudes	Validated	Hybrid x X	0.454	0.176	0.010	0.039	0.008
F2	Pol. role	Generalist	AI x X	-0.052	0.215	0.808	1.000	1.000
F2	Pol. role	Generalist	Hybrid x X	0.060	0.169	0.725	1.000	1.000
F2	Pol. role	Validated	AI x X	0.083	0.191	0.664	1.000	1.000
F2	Pol. role	Validated	Hybrid x X	-0.040	0.176	0.821	1.000	1.000
F2	Municipality $<$ 20k	Generalist	AI x X	0.440	0.216	0.041	1.000	0.231
F2	Municipality $<$ 20k	Generalist	Hybrid x X	-0.287	0.170	0.092	1.000	0.346
F2	Municipality $<$ 20k	Validated	AI x X	0.482	0.193	0.012	0.421	0.123
F2	Municipality $<$ 20k	Validated	Hybrid x X	-0.094	0.177	0.595	1.000	1.000
F2	North	Generalist	AI x X	-0.370	0.255	0.147	1.000	0.399
F2	North	Generalist	Hybrid x X	0.128	0.194	0.510	1.000	1.000
F2	North	Validated	AI x X	0.167	0.216	0.439	1.000	1.000
F2	North	Validated	Hybrid x X	-0.019	0.200	0.925	1.000	1.000
F2	Female	Generalist	AI x X	-0.151	0.219	0.489	1.000	1.000
F2	Female	Generalist	Hybrid x X	0.106	0.179	0.554	1.000	1.000
F2	Female	Validated	AI x X	-0.041	0.193	0.833	1.000	1.000
F2	Female	Validated	Hybrid x X	0.101	0.180	0.574	1.000	1.000
F2	Age $\geq$ 56	Generalist	AI x X	-0.018	0.220	0.933	1.000	1.000
F2	Age $\geq$ 56	Generalist	Hybrid x X	0.008	0.173	0.964	1.000	1.000
F2	Age $\geq$ 56	Validated	AI x X	-0.077	0.193	0.688	1.000	1.000
F2	Age $\geq$ 56	Validated	Hybrid x X	0.143	0.177	0.418	1.000	1.000
F2	Uni. +	Generalist	AI x X	-0.577	0.232	0.013	0.423	0.123
F2	Uni. +	Generalist	Hybrid x X	0.063	0.195	0.747	1.000	1.000
F2	Uni. +	Validated	AI x X	-0.333	0.219	0.127	1.000	0.399
F2	Uni. +	Validated	Hybrid x X	0.162	0.188	0.390	1.000	1.000
F2	AI use	Generalist	AI x X	0.342	0.219	0.118	1.000	0.399
F2	AI use	Generalist	Hybrid x X	0.612	0.169	0.000	0.011	0.011

Continued on next page

Table A18: Heterogeneity interaction coefficients with multiple-testing adjustment (Continued)

Family	Dimension	Arm	Term	Estimate	SE	p	Holm p	Sharpened q
F2	AI use	Validated	AI x X	0.611	0.192	0.001	0.051	0.027
F2	AI use	Validated	Hybrid x X	0.393	0.176	0.026	0.828	0.199
F2	Science use	Generalist	AI x X	-0.014	0.231	0.952	1.000	1.000
F2	Science use	Generalist	Hybrid x X	0.187	0.179	0.297	1.000	0.772
F2	Science use	Validated	AI x X	-0.305	0.211	0.148	1.000	0.399
F2	Science use	Validated	Hybrid x X	0.380	0.193	0.048	1.000	0.231
F2	Science useful	Generalist	AI x X	-0.683	0.344	0.047	1.000	0.231
F2	Science useful	Generalist	Hybrid x X	0.291	0.294	0.322	1.000	0.794
F2	Science useful	Validated	AI x X	-0.619	0.335	0.065	1.000	0.278
F2	Science useful	Validated	Hybrid x X	0.100	0.340	0.768	1.000	1.000
F2b	Economy and labour	Generalist	AI x X	0.194	0.225	0.390	1.000	1.000
F2b	Economy and labour	Generalist	Hybrid x X	0.121	0.189	0.522	1.000	1.000
F2b	Economy and labour	Validated	AI x X	0.367	0.198	0.064	1.000	0.721
F2b	Economy and labour	Validated	Hybrid x X	0.333	0.184	0.070	1.000	0.721
F2b	Education and research	Generalist	AI x X	0.267	0.275	0.331	1.000	1.000
F2b	Education and research	Generalist	Hybrid x X	0.087	0.200	0.662	1.000	1.000
F2b	Education and research	Validated	AI x X	-0.142	0.242	0.558	1.000	1.000
F2b	Education and research	Validated	Hybrid x X	0.118	0.220	0.591	1.000	1.000
F2b	Health	Generalist	AI x X	0.040	0.293	0.892	1.000	1.000
F2b	Health	Generalist	Hybrid x X	-0.037	0.235	0.875	1.000	1.000
F2b	Health	Validated	AI x X	0.759	0.297	0.011	0.327	0.203
F2b	Health	Validated	Hybrid x X	0.841	0.326	0.010	0.320	0.203
F2b	Social policies	Generalist	AI x X	0.151	0.229	0.511	1.000	1.000
F2b	Social policies	Generalist	Hybrid x X	-0.155	0.181	0.392	1.000	1.000
F2b	Social policies	Validated	AI x X	0.185	0.201	0.356	1.000	1.000
F2b	Social policies	Validated	Hybrid x X	0.231	0.192	0.229	1.000	1.000
F2b	Environment and sustainability	Generalist	AI x X	-0.346	0.235	0.140	1.000	1.000
F2b	Environment and sustainability	Generalist	Hybrid x X	0.010	0.182	0.955	1.000	1.000
F2b	Environment and sustainability	Validated	AI x X	-0.382	0.211	0.071	1.000	0.721
F2b	Environment and sustainability	Validated	Hybrid x X	-0.104	0.188	0.578	1.000	1.000
F2b	Security, justice, foreign policy, defence	Generalist	AI x X	0.186	0.355	0.600	1.000	1.000
F2b	Security, justice, foreign policy, defence	Generalist	Hybrid x X	-0.040	0.246	0.870	1.000	1.000

Continued on next page

Table A18: Heterogeneity interaction coefficients with multiple-testing adjustment (Continued)

Family	Dimension	Arm	Term	Estimate	SE	p	Holm p	Sharpened q
F2b	Security, justice, foreign policy, defence	Validated	AI x X	-0.588	0.340	0.084	1.000	0.721
F2b	Security, justice, foreign policy, defence	Validated	Hybrid x X	0.316	0.287	0.271	1.000	1.000
F2b	Infrastructure and transport	Generalist	AI x X	-0.294	0.253	0.246	1.000	1.000
F2b	Infrastructure and transport	Generalist	Hybrid x X	0.261	0.185	0.159	1.000	1.000
F2b	Infrastructure and transport	Validated	AI x X	-0.085	0.229	0.711	1.000	1.000
F2b	Infrastructure and transport	Validated	Hybrid x X	-0.199	0.205	0.331	1.000	1.000
F2b	Culture, tourism and sport	Generalist	AI x X	-0.034	0.223	0.877	1.000	1.000
F2b	Culture, tourism and sport	Generalist	Hybrid x X	0.056	0.177	0.751	1.000	1.000
F2b	Culture, tourism and sport	Validated	AI x X	-0.254	0.204	0.213	1.000	1.000
F2b	Culture, tourism and sport	Validated	Hybrid x X	-0.136	0.182	0.455	1.000	1.000
F3	Decision-maker	Generalist	AI x X	-0.224	0.248	0.366	1.000	1.000
F3	Decision-maker	Generalist	Hybrid x X	-0.153	0.210	0.468	1.000	1.000
F3	Decision-maker	Validated	AI x X	-0.065	0.221	0.768	1.000	1.000
F3	Decision-maker	Validated	Hybrid x X	-0.006	0.198	0.976	1.000	1.000
F3	Analyst	Generalist	AI x X	0.400	0.233	0.085	1.000	1.000
F3	Analyst	Generalist	Hybrid x X	0.184	0.176	0.295	1.000	1.000
F3	Analyst	Validated	AI x X	0.077	0.204	0.707	1.000	1.000
F3	Analyst	Validated	Hybrid x X	0.250	0.184	0.175	1.000	1.000
F3	Tasks > median	Generalist	AI x X	0.016	0.215	0.942	1.000	1.000
F3	Tasks > median	Generalist	Hybrid x X	-0.057	0.171	0.738	1.000	1.000
F3	Tasks > median	Validated	AI x X	0.039	0.192	0.840	1.000	1.000
F3	Tasks > median	Validated	Hybrid x X	0.204	0.178	0.251	1.000	1.000

Notes: Interaction coefficients from multinomial logit models estimated separately by arm (Generalist AI and Validated AI), with cluster-robust standard errors. AI × X and Hybrid × X are the interactions of Automatic AI and Hybrid with the dimension indicator. Families: F1, main heterogeneity dimensions (attitudes towards AI and science); F2, other pre-registered dimensions; F2b, policy areas (pre-registered); F3, exploratory (policy tasks). Holm-adjusted p-values and Anderson (2008) sharpened q-values are computed within family.

Table A19: Policy tasks performed in the previous year, by role

	Executive (N=134)	Legislative (N=269)	Admin. Director (N=277)	Admin. Official (N=98)	Other (N=35)	All (N=813)
Study possible solutions to problems requiring public intervention	0.96	0.84	0.90	0.77	0.69	0.86
Decide which solutions to adopt	0.95	0.62	0.87	0.73	0.43	0.77
Translate solutions into regulatory and operational guidance	0.70	0.40	0.89	0.70	0.40	0.65
Produce reports / structured information supporting policies	0.37	0.25	0.82	0.57	0.46	0.51
Monitor the implementation of solutions	0.84	0.65	0.89	0.65	0.60	0.76
Design or conduct impact evaluations	0.39	0.27	0.47	0.24	0.43	0.36
Conduct or supervise statistical analyses	0.25	0.16	0.41	0.26	0.23	0.28
None of the seven tasks	0.02	0.06	0.02	0.08	0.23	0.05
Mean number of tasks (0-7)	4.45	3.20	5.24	3.93	3.23	4.19
Named a topic for a commissioned synthesis (Q14)	0.81	0.78	0.78	0.59	0.37	0.74
Lack of evidence syntheses is a barrier	0.28	0.35	0.39	0.23	0.20	0.33

Notes: Shares of respondents answering Yes, analytic sample (N = 813). Admin. Director: director/manager; Admin. Official: official/civil servant; Other: other employees, consultants, party officials and other roles. Q14 is an open-ended question asked before the DCE; a respondent is counted if they entered any text.

Table A20: Heterogeneity analysis by policy tasks performed (exploratory)

	Generalist	Validated	Generalist	Validated	Generalist	Validated
Automatic AI	-2.089***	-1.258***	-2.521***	-1.363***	-2.267***	-1.325***
	(0.231)	(0.200)	(0.219)	(0.181)	(0.173)	(0.151)
Hybrid (AI + Expert)	0.388*	0.715***	0.152	0.547***	0.296*	0.619***
	(0.192)	(0.174)	(0.147)	(0.155)	(0.120)	(0.124)
AI $\times X$	-0.224	-0.065	0.400	0.077	0.016	0.039
	(0.248)	(0.221)	(0.233)	(0.204)	(0.215)	(0.192)
Hybrid $\times X$	-0.153	-0.006	0.184	0.250	-0.057	0.204
	(0.210)	(0.198)	(0.176)	(0.184)	(0.171)	(0.178)
Cost (€000)	-0.277***	-0.253***	-0.276***	-0.252***	-0.277***	-0.250***
	(0.042)	(0.041)	(0.042)	(0.040)	(0.042)	(0.041)
Time (weeks)	-0.082***	-0.082***	-0.082***	-0.083***	-0.082***	-0.082***
	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)
Log-Likelihood	-1564.3	-1652.1	-1561.2	-1650.2	-1565.4	-1650.8
$X =$	Decision-maker	Decision-maker	Analyst	Analyst	Tasks > median	Tasks > median

Notes: Coefficients estimated via multinomial logit. Clustered standard errors at respondent level in parentheses. The reference category for the synthesis method is the "Expert" produced synthesis status quo. The "Automatic AI" option refers to synthesis produced by a general-purpose model (*Generalist AI*) or a validated model (*Validated AI*). Total $N = 813$ respondents $\times$ 8 choice tasks $\times$ 2 alternatives Asterisks indicate significance at the *** 0.1%, ** 1%, and * 5% levels.

Table A21: Robustness: excluding respondents with no reported policy task or no commissioning topic

	Excl. respondents reporting no task			Excl. respondents naming no topic (Q14)		
	Full	Generalist	Validated	Full	Generalist	Validated
Automatic AI	-1.774***	-2.277***	-1.322***	-1.797***	-2.293***	-1.365***
	(0.094)	(0.141)	(0.126)	(0.106)	(0.166)	(0.137)
Hybrid (AI + Expert)	0.496***	0.271**	0.728***	0.502***	0.366***	0.638***
	(0.069)	(0.094)	(0.101)	(0.079)	(0.109)	(0.115)
Cost (€000)	-0.266***	-0.266***	-0.272***	-0.252***	-0.242***	-0.271***
	(0.029)	(0.043)	(0.042)	(0.033)	(0.048)	(0.046)
Time (weeks)	-0.084***	-0.088***	-0.082***	-0.088***	-0.085***	-0.095***
	(0.013)	(0.019)	(0.020)	(0.015)	(0.023)	(0.022)
N obs.	12352	6464	5888	9664	5120	4544
Log-Likelihood	-3099.6	-1487.5	-1555.1	-2409.8	-1142.1	-1218.4

Notes: Coefficients estimated via multinomial logit. Clustered standard errors at respondent level in parentheses. Reference category: "Expert" produced synthesis. Asterisks indicate significance at the *** 0.1%, ** 1%, and * 5% levels.

Table A22: Relative attribute importance

Attribute	Full sample	Generalist AI	Validated AI
Production method	0.58	0.59	0.55
Cost	0.25	0.24	0.26
Delivery time	0.17	0.16	0.19

Notes: Utility range of each attribute over its experimental levels divided by the sum of the ranges, from the MNL estimates in Table 3.

Table A23: Heterogeneity analysis by geographic area: North vs. Centre, South and Islands

	Generalist	Validated
Automatic AI	-1.986***	-1.441***
	(0.240)	(0.205)
Hybrid (AI + Expert)	0.172	0.715***
	(0.170)	(0.179)
AI $\times X$	-0.370	0.167
	(0.255)	(0.216)
Hybrid $\times X$	0.128	-0.019
	(0.194)	(0.200)
Cost (€000)	-0.277***	-0.254***
	(0.042)	(0.041)
Time (weeks)	-0.082***	-0.083***
	(0.019)	(0.019)
Log-Likelihood	-1546.4	-1646.3
$X =$	North	North

Notes: Coefficients estimated via multinomial logit. Clustered standard errors at respondent level in parentheses. The reference category for the synthesis method is the "Expert" produced synthesis status quo. The "Automatic AI" option refers to synthesis produced by a general-purpose model (*Generalist AI*) or a validated model (*Validated AI*). Total $N = 813$ respondents $\times$ 8 choice tasks $\times$ 2 alternatives Asterisks indicate significance at the *** 0.1%, ** 1%, and * 5% levels.

Table A24: Heterogeneity by geographic area, three groups (supplementary, low power)

	Generalist	Validated
Automatic AI	-2.353***	-1.273***
	(0.150)	(0.132)
Hybrid (AI + Expert)	0.302**	0.697***
	(0.106)	(0.109)
AI $\times$ Centre	0.148	-0.260
	(0.313)	(0.269)
Hybrid $\times$ Centre	-0.268	0.027
	(0.228)	(0.244)
AI $\times$ South & Islands	0.640	-0.025
	(0.369)	(0.307)
Hybrid $\times$ South & Islands	0.064	0.002
	(0.300)	(0.295)
Cost (€000)	-0.277***	-0.254***
	(0.042)	(0.041)
Time (weeks)	-0.081***	-0.082***
	(0.019)	(0.019)
Log-Likelihood	-1544.3	-1645.8

Notes: Reference: North. Clustered standard errors at respondent level in parentheses. Asterisks indicate significance at the *** 0.1%, ** 1%, and * 5% levels.

Table A25: Latent class logit: three-class model

	Class 1	Class 2	Class 3
Automatic AI	-0.613*** (0.141)	-4.341*** (0.337)	-3.817*** (0.306)
Hybrid (AI + Expert)	1.249*** (0.114)	-1.601*** (0.208)	1.152*** (0.157)
Cost (€000)	-0.715*** (0.063)	-0.171* (0.083)	-0.388*** (0.079)
Time (weeks)	-0.197*** (0.028)	-0.038 (0.039)	-0.170*** (0.038)
Membership: constant	(ref.)	0.171 (0.203)	0.883*** (0.174)
Membership: Validated AI	(ref.)	-1.193*** (0.248)	-1.023*** (0.203)
Membership: AI attitudes (std.)	(ref.)	-1.332*** (0.148)	-0.360** (0.118)
Class share (all)	0.33	0.25	0.42
Class share (Generalist AI)	0.22	0.30	0.48
Class share (Validated AI)	0.44	0.20	0.36

Notes: Panel latent class logit estimated by maximum likelihood; standard errors from the numerical Hessian in parentheses. Class membership is a multinomial logit in the Validated AI indicator and standardised AI attitudes (Class 1 is the reference). Class shares are average predicted prior membership probabilities. Information criteria for 1-4 classes (BIC): 6576, 5699, 5440, 5370. Asterisks indicate significance at the *** 0.1%, ** 1%, and * 5% levels.

Table A26: Hybrid premium by choice-set composition

	Full Sample	Generalist AI	Validated AI
Automatic AI	-1.520***	-2.022***	-1.079***
	(0.091)	(0.138)	(0.122)
Hybrid, in Expert-vs-Hybrid tasks	0.307***	0.145	0.468***
	(0.070)	(0.097)	(0.101)
Hybrid, in Hybrid-vs-AI tasks	1.151***	0.891***	1.385***
	(0.111)	(0.154)	(0.156)
Cost (€000)	-0.256***	-0.269***	-0.254***
	(0.028)	(0.041)	(0.040)
Time (weeks)	-0.077***	-0.077***	-0.081***
	(0.012)	(0.018)	(0.018)
Log-Likelihood	-3238.3	-1553.0	-1629.5
p-value: equal Hybrid coefficients	0.000	0.000	0.000

Notes: Coefficients estimated via multinomial logit. Clustered standard errors at respondent level in parentheses. Reference category: "Expert" produced synthesis. The Hybrid coefficient is allowed to differ between tasks in which Hybrid is compared with Expert and tasks in which it is compared with the Automatic AI option. The last row reports the p-value of a Wald test of equality of the two Hybrid coefficients. Asterisks indicate significance at the *** 0.1%, ** 1%, and * 5% levels.

Table A27: Distribution of individual-level conditional mean coefficients (mixed logit)

	Generalist AI	Validated AI
AI coef. > 0	0.02	0.04
Hybrid coef. > 0	0.61	0.84
Hybrid > 0 and AI > 0	0.00	0.04
Hybrid < 0 (prefer Expert)	0.39	0.16
Corr(AI coef., AI attitudes)	0.14	0.26
Corr(Hybrid coef., AI attitudes)	0.32	0.29

Notes: Shares and correlations computed from the respondent-level conditional means of the random coefficients of the mixed logit models in Table 6. Conditional means shrink towards the population mean and understate the spread of the underlying distribution.

Table A28: Willingness-to-pay estimates and the support of the experimental design

Sample	Attribute	WTP	CI low	CI high	Max. cost difference in design	Inside design support
Full sample	Automatic AI (vs. Expert)	-6758	-8116	-5399	3750	No
Full sample	Hybrid (vs. Expert)	1874	1174	2574	2000	Yes
Generalist AI	Automatic AI (vs. Expert)	-8155	-10415	-5895	3750	No
Generalist AI	Hybrid (vs. Expert)	970	207	1734	2000	Yes
Validated AI	Automatic AI (vs. Expert)	-5186	-6728	-3644	3750	No
Validated AI	Hybrid (vs. Expert)	2816	1564	4068	2000	No

Notes: WTP in euros from Table A7. The maximum cost difference is the largest price gap between the two alternatives that respondents could face in the relevant pair type (Expert minus AI; Hybrid minus Expert). Estimates outside the support are extrapolations and should be read as indicating strong relative preferences rather than amounts respondents would pay.

Table A29: Predicted probability of choosing alternative A (MNL)

Scenario	Full sample	Generalist AI	Validated AI
Expert (EUR 4,000, 2 months) vs. AI (EUR 250, 1 day)	0.52	0.63	0.42
Expert (EUR 2,000, 1 month) vs. AI (EUR 1,000, 1 week)	0.77	0.85	0.69
Expert (EUR 4,000, 2 months) vs. Hybrid (EUR 2,000, 1 month)	0.20	0.24	0.17
Expert vs. Hybrid, same cost and time (EUR 2,000, 1 month)	0.38	0.43	0.33
Hybrid (EUR 2,000, 1 month) vs. AI (EUR 500, 1 week)	0.83	0.86	0.80

Notes: Predicted probabilities from the multinomial logit models in Table 3. All attribute levels are within the experimental design. Each entry is the predicted probability of choosing the first-listed alternative in the scenario (alternative A); the probability of choosing the second alternative is its complement. For example, in the first scenario the probability of choosing the AI synthesis is $1 - 0.42 = 0.58$ in the Validated AI condition and $1 - 0.63 = 0.37$ in the Generalist AI condition.

Table A30: Main estimates among respondents with less positive attitudes towards AI or low GenAI use

	AI attitudes below median			Never / rarely use GenAI		
	Full	Generalist	Validated	Full	Generalist	Validated
Automatic AI	-2.160***	-2.531***	-1.823***	-1.908***	-2.197***	-1.668***
	(0.142)	(0.212)	(0.191)	(0.133)	(0.208)	(0.175)
Hybrid (AI + Expert)	0.151	-0.028	0.334*	0.290**	0.089	0.492***
	(0.099)	(0.137)	(0.144)	(0.093)	(0.134)	(0.131)
Cost (€000)	-0.213***	-0.174**	-0.252***	-0.227***	-0.186***	-0.259***
	(0.042)	(0.057)	(0.061)	(0.041)	(0.056)	(0.060)
Time (weeks)	-0.063***	-0.070**	-0.061*	-0.073***	-0.052	-0.095***
	(0.019)	(0.026)	(0.027)	(0.018)	(0.027)	(0.026)
N obs.	6512	3520	2992	6432	3232	3200
Log-Likelihood	-1562.7	-779.5	-764.8	-1617.5	-764.1	-828.9

Notes: Coefficients estimated via multinomial logit. Clustered standard errors at respondent level in parentheses. Reference category: "Expert" produced synthesis. Asterisks indicate significance at the *** 0.1%, ** 1%, and * 5% levels.

Survey materials

Survey invitation e-mail

Subject: Invitation to participate in the {Project Name} questionnaire

{Institute} {Project Name} <{Project E-mail}> 22 July 2025 at 08:00
Reply to: <{Project E-mail}>
To: {Recipient E-mail}

Dear {Recipient Name},

We are contacting you on behalf of {Name}, Director of {Institute} in {City}, to invite you to participate in a brief questionnaire as part of the {Project Name} project, funded by {Funding Programme}.

The project aims to develop a platform that provides public administrations and political bodies with summaries of scientific literature on the impact of public policies, in order to support evidence-based decision-making.

The questionnaire only takes a few minutes and will help us calibrate the tool to the real needs of those who contribute to decisions of public relevance.

Link to the questionnaire: {Link}

Attached you will find the official letter from {Name} with further details about the project.

For further information:

- Email: {Project E-mail}
- Contact person: {Researcher Name} ({E-mail} –{Phone})
- Website: dedicated page on {Website}

We thank you in advance for your valuable collaboration.

Kind regards,

{Institute}

Attachment: Invite.pdf (241K)

Survey invitation letter

{Name}
{Institute}
{Address}
{E-mail}

{City}, 18 July 2025

Subject: Participation in a short questionnaire for the {Project Name} project funded by {Funding Programme}

Dear {Recipient Name},

I am writing to you as Director of {Institute} in {City}. {Institute} has collaborated for many years with various institutions and public bodies both in evaluating the effects of policies and in building a knowledge base aimed at supporting policymakers in the design of those policies.

{Institute} is currently working on the {Project Name} project, funded by {Funding Programme}. The project aims to develop a platform that makes available to public administrations and political bodies summaries of the scientific literature on the impact of public policies across various fields.

The project adopts a collaborative approach, emphasising the importance of ongoing dialogue between the research sector and public institutions with a view to continuous innovation and improvement.

I am therefore writing to ask whether you would be willing to complete the short questionnaire available at the link below, which is designed to gauge interest in the use of scientific evidence and the topics of relevance to the organisation in which you work.

The questionnaire takes only a few minutes to complete and represents a valuable contribution towards developing a tool that is genuinely tailored to the needs of those who daily contribute to decisions of public significance.

The questionnaire is available at the following address: {Link}.

I remain at your disposal for any further information. For additional details or technical clarifications, please feel free to send an email to {Project E-mail} or contact researcher {Name} directly ({E-mail} – {Phone}), or visit the dedicated page on our institutional website.

Thanking you in advance for your time and cooperation, I take this opportunity to extend my warmest regards.

{Name}
{Signature}

Survey questions

Role

Q1. Which of the following categories best describes your role in public administration and/or public policy?

a) Political role (e.g., assessor, mayor, councillor, member of parliament, etc.)
b) Administrative role (e.g., manager, director, official, employee, consultant, etc.)
c) I do not have a role in public administration and/or public policy *[exit]*

Q2. *[If $Q1 = a$]*

If you hold a political role, which of the following categories best describes your main role?

a) Member of an executive body (e.g., President of a Regional Council, Assessor, Mayor, Government Member, etc.)
b) Member of a legislative/assembly body (e.g., Councillor, President of a Council, Member of Parliament, etc.)
c) Member of a political party, movement, or trade union in an active role (e.g., Party Secretary, Provincial Coordinator, Secretary-General, etc.). *Select this option only if you are not also a member of an executive or legislative/assembly body.*
d) Other

Q3. *[If $Q1 = b$]*

If you hold an administrative role, which of the following categories best describes your level?

a) Manager / Director / Senior Officer
b) Official
c) Other employee
d) Consultant
e) Other

Q4. What territorial level does your organisation belong to?

a) Municipal – fewer than 2,000 inhabitants
b) Municipal – between 2,000 and 20,000 inhabitants
c) Municipal – more than 20,000 inhabitants
d) Provincial or Metropolitan City
e) Regional
f) National or European
g) Other

Q5. *[If $Q4 \neq f$]*

What is the geographic area of your organisation?

a) North-East

b) North-West

c) Centre

d) South

e) Islands

Activities related to public intervention

Q6. In the past year, have you carried out any of the following activities?

[Yes/No answer for each activity]

a) Studying possible solutions to problems requiring public intervention

b) Deciding which solutions to adopt

c) Translating solutions into regulatory and operational guidelines

d) Producing reports and other types of structured information to support public policies

e) Monitoring the implementation of solutions

f) Designing or conducting impact evaluations of public policies or interventions

g) Conducting or supervising statistical analyses to support public decisions

Q7. What is/are your area(s) of expertise?

a) Economics and Labour

b) Education and Research

c) Health / Healthcare

d) Social Policy

e) Environment and Sustainability

f) Security, Justice, Foreign Policy and Defence

g) Infrastructure and Transport

h) Culture, Tourism and Sport

Background

Q8. You are:

a) Woman

b) Man

c) I do not identify with either category

d) Prefer not to answer

Q9. Age group:

a) 18 to 35 years old

b) 36 to 55 years old

c) 56 to 65 years old

d) Over 65 years old

e) Prefer not to answer

Q10. Please indicate your highest level of education:

a) Secondary school diploma or lower
b) Bachelor's degree
c) Postgraduate qualification
d) Prefer not to answer

Use of scientific studies (random order with next section)

Q11. How often do you consult (directly, or indirectly through colleagues or consultants) the findings of scientific studies?
By 'scientific studies' we mean articles published in peer-reviewed academic journals.

a) Never
b) Rarely
c) Sometimes
d) Often
e) Always

Q12. In your opinion, which of the following factors limit the use of scientific studies in the design of public policy interventions?

a) Lack of studies applicable to the specific context
b) Overly technical language
c) Difficulty in translating studies into concrete action
d) Lack of dedicated staff
e) Difficulty in identifying or accessing studies
f) Lack of financial resources
g) Language barrier (studies in a foreign language)
h) Lack of time
i) Greater trust in direct field experience
j) Lack of synthetic reviews of scientific studies
k) Concerns about the quality and reliability of studies

Q13. How useful do you consider the findings of scientific studies in the design of public policy interventions?

a) Not at all useful
b) Slightly useful
c) Fairly useful
d) Very useful

Q14. If you could commission a synthesis of scientific findings, on which topics would you request it?
E.g., policies to address falling birth rates, incentive policies for household waste recycling, regulation of the short-term rental market, the impact of unemployment benefits on labour market re-entry, etc.

Please indicate the extent to which you agree or disagree with the following statements. There are no right or wrong answers.

Q15. Scientific research methods are the best way to discover whether something is true or false.

a) Strongly disagree
b) Disagree
c) Neither agree nor disagree
d) Agree
e) Strongly agree

Q16. Scientists should work closely with policymakers to integrate scientific findings into decision-making.

a) Strongly disagree
b) Disagree
c) Neither agree nor disagree
d) Agree
e) Strongly agree

Q17. Scientists should be more involved in policy-making.

a) Strongly disagree
b) Disagree
c) Neither agree nor disagree
d) Agree
e) Strongly agree

Q18. Scientists should communicate their findings to policymakers.

a) Strongly disagree
b) Disagree
c) Neither agree nor disagree
d) Agree
e) Strongly agree

Use of artificial intelligence (random order with previous section)

Q19. How often do you use generative Artificial Intelligence models (e.g., ChatGPT, Claude, Gemini, etc.)?

a) Never
b) Rarely

c) Sometimes

d) Often

e) Always

Please indicate the extent to which you agree or disagree with the following statements. There are no right or wrong answers.

Q20. I have positive feelings when I think about using artificial intelligence in my daily work life.

a) Strongly disagree

b) Disagree

c) Neither agree nor disagree

d) Agree

e) Strongly agree

Q21. At work, artificial intelligence offers valid solutions for many tasks.

a) Strongly disagree

b) Disagree

c) Neither agree nor disagree

d) Agree

e) Strongly agree

Q22. If I had to complete an important task at work, I would prefer to use a technology with artificial intelligence rather than one without.

a) Strongly disagree

b) Disagree

c) Neither agree nor disagree

d) Agree

e) Strongly agree

Experimental instructions

Before completing the discrete choice experiment, respondents were presented with general instructions and a hypothetical policy-making scenario (see Figures A1 and A2). They were asked to imagine that they had to develop, within three months, a new public policy proposal addressing a particularly relevant issue within their area of competence. To support this task, they could commission a synthesis of scientific studies evaluating the impact of public policies related to the issue at hand. Respondents were then informed that, in each of the following eight choice tasks, they would be asked to choose between two alternative options for producing this synthesis. The alternatives varied along three attributes: who produces the synthesis and with what approach, delivery time, and cost.

Figure A1: General instructions

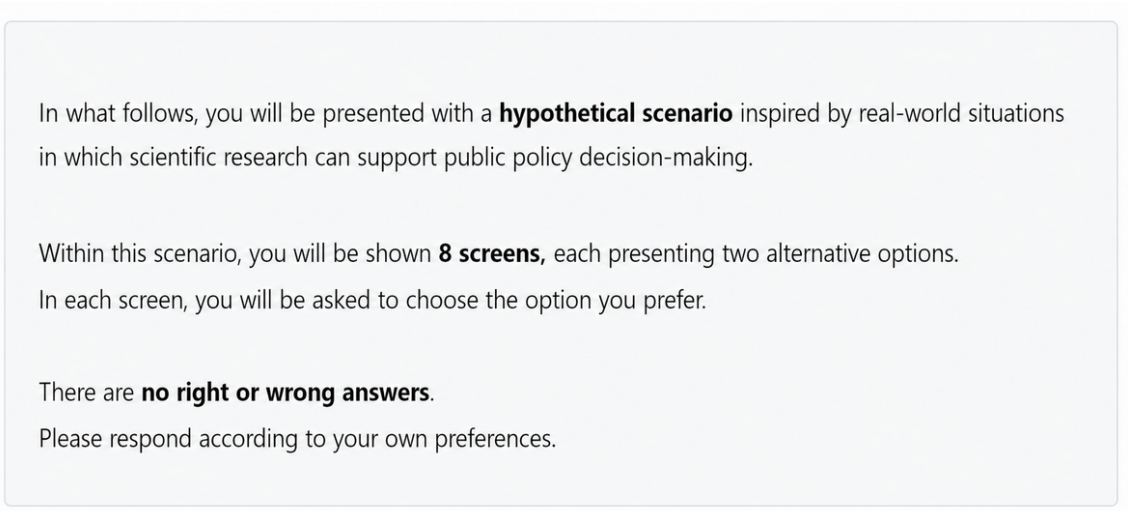
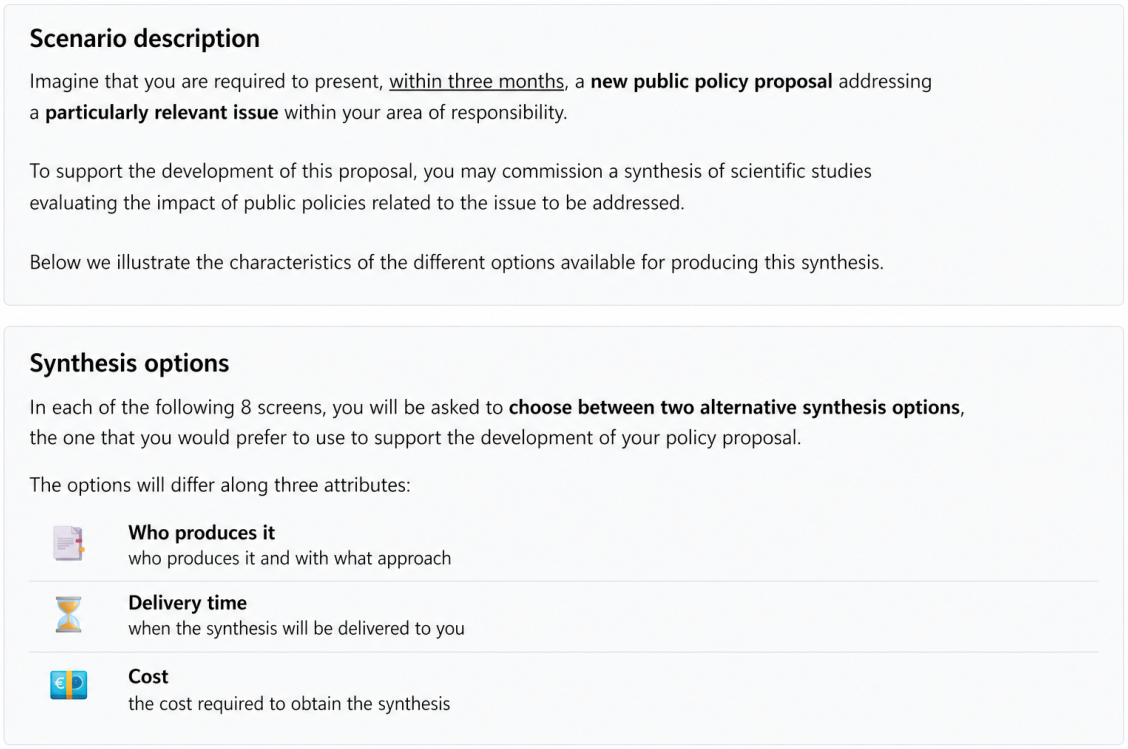


Figure A2: Scenario description



Respondents received detailed instructions on the meaning of the production-method attribute before completing the choice tasks. Figures A3 and A4 report the attribute description shown for each of the two treatment arms: the validated AI and general-purpose AI conditions, respectively.

Figure A3: Attributes description in the validated AI condition

Who produces it

The attribute **who produces it** can take one of the following three values:

Experts

Experts (academics or consultants with relevant expertise) produce a document that synthesises the scientific evidence.

Validated AI

A **generative** artificial intelligence (AI) model **validated** for the synthesis of scientific studies automatically produces a document that synthesises the scientific evidence.

Validated AI with expert review

A **generative** artificial intelligence (AI) model, **validated** for the synthesis of scientific studies, automatically produces a document that synthesises the scientific evidence, which is then **reviewed by experts**.

What does validated AI mean?

It refers to a generative AI that has been specifically tested in the synthesis of scientific studies, by comparing the results produced with those of human experts.

The following figure represents in a simplified way the **process of creating** the three types of synthesis. In all cases, the starting point is the same: a common set of scientific studies relevant to the issue to be addressed.
The **difference** lies in who analyses these studies and in which way.

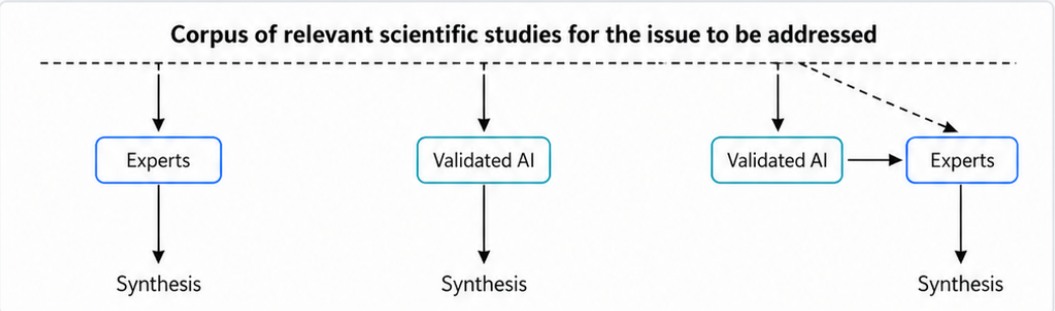
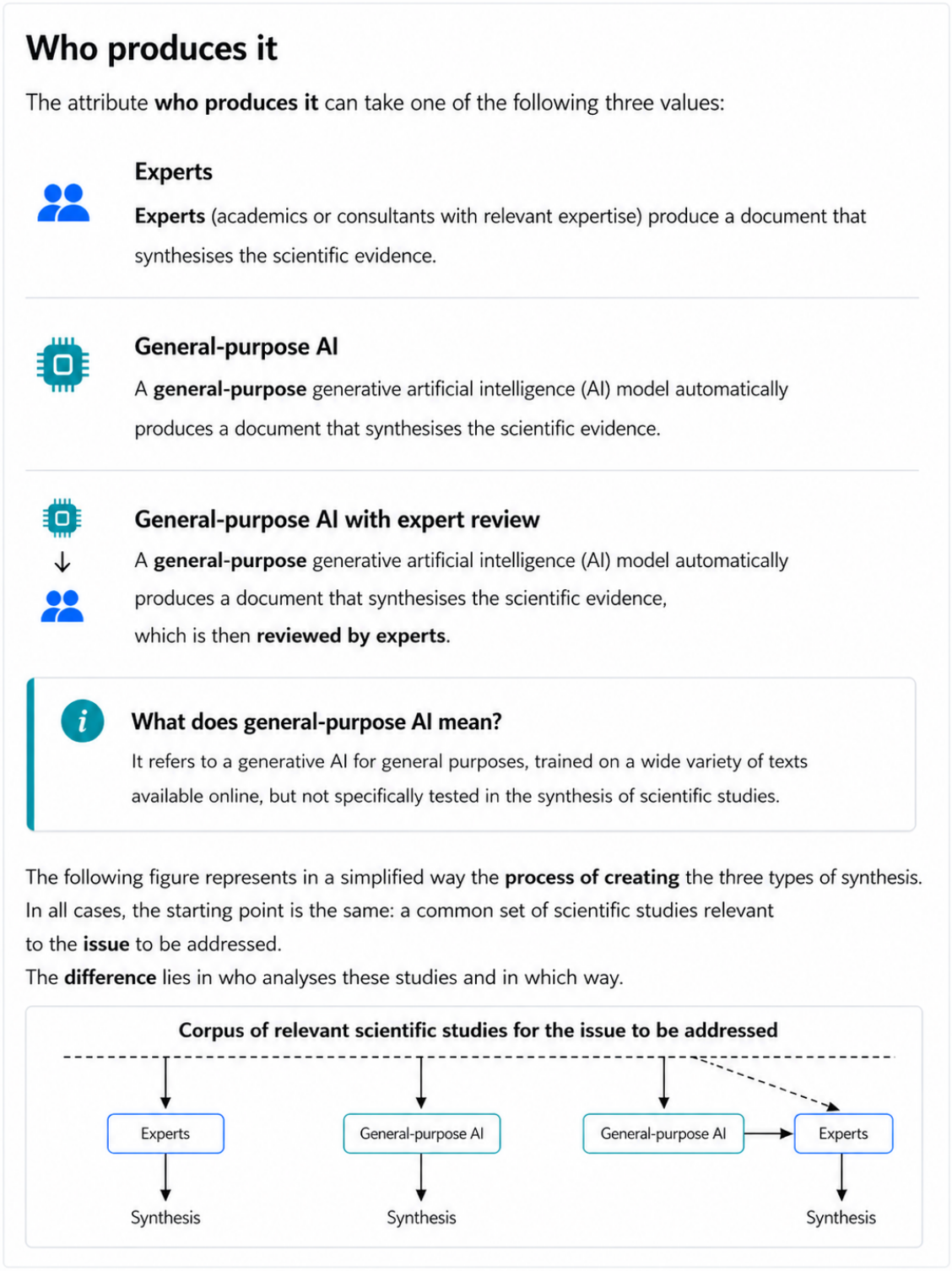


Figure A4: Attributes description in the general-purpose AI condition



Each choice task presented respondents with two alternative synthesis options displayed side by side (see Figures A5 and A6). Respondents were reminded of the policy-making scenario and asked to choose the synthesis they would prefer to support the development of the policy proposal.

The alternatives differed along three attributes: the production method, delivery time, and cost. In the example shown, respondents choose between a synthesis produced by a validated AI system with expert review and one produced by validated AI alone, with the two options also varying in delivery time and cost. Respondents indicated their preferred option by selecting one of the two alternatives.

Figure A5: Example choice task: validated AI condition

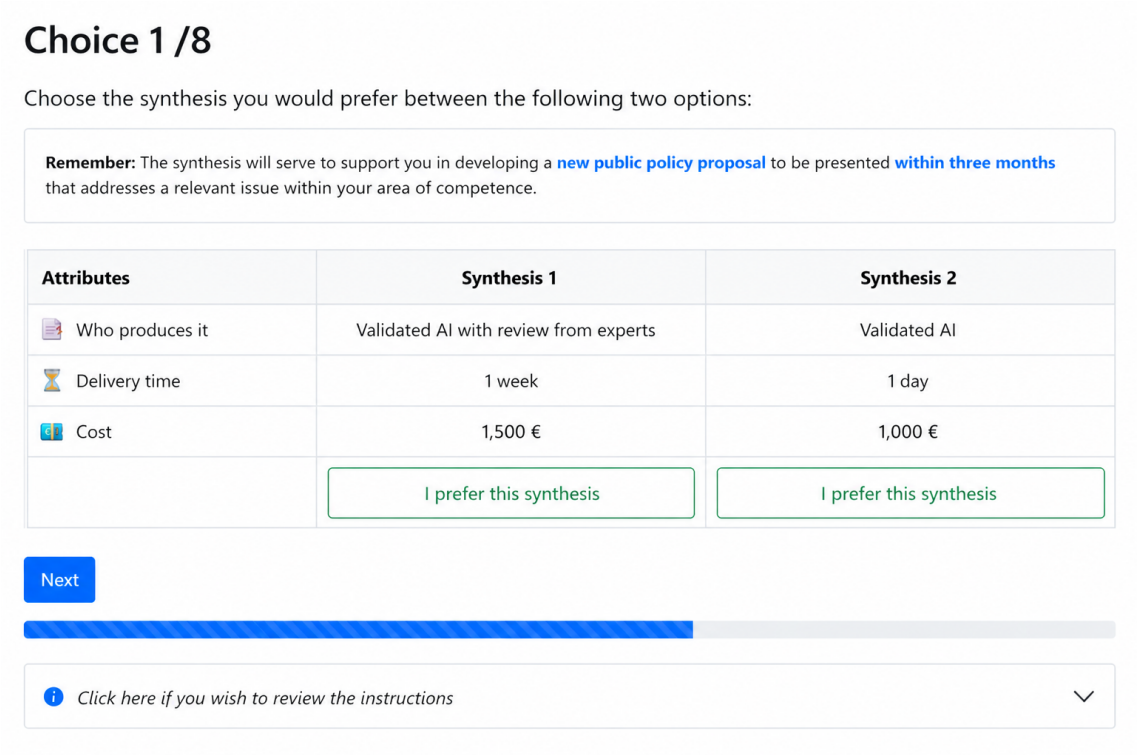


Figure A6: Example choice task: general-purpose AI condition

