Innovation Growth Lab

POLICY BRIEF

Building evidence on AI-assisted review in public funding

The Neotec shadow experiment with CDTI

IGL Policy Brief · September 2026

Hugo Cuello
Senior Policy Analyst

Executive Summary

This policy brief presents how the Innovation Growth Lab (IGL) supported the Spanish Innovation Agency (CDTI) in designing and implementing a large-scale "shadow experiment" within the Neotec programme, one of Spain's flagship funding schemes for high-tech start-ups. The project emerged in response to growing concerns about the difficulty of assessing social impact within innovation programmes, as well as broader interest in whether AI can help improve public sector decision-making. The experiment created a parallel evaluation environment where alternative assessment approaches, including AI-assisted tools, could be tested safely under real-world conditions.

The results highlight both the potential and the limits of AI in innovation policy. Predicting the future social impact of early-stage innovation projects proved difficult across all evaluator groups, reflecting the high levels of uncertainty and measurement challenges associated with social impact assessment. AI-generated information produced small but consistent improvements in evaluators' predictive capacity, although its effects remained modest and highly dependent on human behaviour and institutional context. More broadly, the project shows that experimentation in government requires not only technical expertise, but also trust, flexibility, institutional support, and the ability to adapt experimental designs to operational realities.

Acknowledgements

This project would not have been possible without the exceptional commitment, motivation, and persistence of María-Ascensión Barajas (CDTI) and Teo Firpo (Humboldt University), whose leadership, optimism, and willingness to navigate the many challenges of the project were fundamental from beginning to end. I am also especially grateful to the Department of Strategy and Evaluation at CDTI, particularly Miguel Valle for his active engagement throughout the project, and Blanca Valbuena for her coordination and organisational support. Special thanks also go to CDTI General Director José Moisés Martín for his forward-looking leadership and openness to experimentation within the organisation, the CDTI Systems Division for helping us with the data, and to all the volunteers who participated in the project. Particular thanks go to David Ampudia for his support and work developing the AI system used in the experiment, and to Hannia González for coordinating the selection of external experts. Earlier conversations with Ina Ganguli (UMASS) and Albert Banal (BSE) were also highly valuable in shaping the project's initial ideas. Finally, I would like to thank BSE for its organisational support and the wider IGL team, particularly James Phipps, Albert Bravo-Biosca, Rob Fuller, Maria Brackin, Sara García Arteagoitia, and others for their feedback, encouragement, and support at different stages of the project.

1. Context

1.1 The Neotec programme

The Centro para el Desarrollo Tecnológico y la Innovación (CDTI-E.P.E.) is Spain's main public agency for supporting business R&D and innovation. It plays a central role in allocating public funding to firms, promoting technological development, and strengthening the national innovation ecosystem.

Within its portfolio, the Neotec programme is a flagship instrument designed to support the creation and consolidation of early-stage technology-based companies. With a budget of approximately €40 million in 2025, it targets firms developing their own R&D-driven technologies, often at low levels of technological maturity, and provides non-repayable grants of up to €250,000 (or €325,000 under specific conditions). The programme operates through competitive calls, where proposals are assessed by CDTI technical evaluators using a structured scoring system that combines innovation, business viability, team capacity, and social impact.

Historically, this scoring system has been heavily oriented towards technological and business criteria. Until recently, social impact accounted for a small share of the total score (around 5%), compared to 35% for technological excellence and 30% each for business viability and team quality. CDTI has taken steps to strengthen this dimension, increasing its weight and revising its definition in recent calls.

The programme has been broadly successful in achieving its core objectives. Existing evaluations point out that Neotec has a clear positive impact on R&D investment, technology transfer, patenting activity, and firm survival. However, these strengths coexist with limitations in other social dimensions, which have become an increasing focus of policy attention.

1.2 The challenge of assessing social impact

When measuring the impacts of Neotec, CDTI found that while the programme performs strongly in technological and innovation-related dimensions, it shows limited effects on social impact. For instance, no significant effects are observed in areas such as gender equality, inclusive innovation, or the development of technologies addressing societal challenges. Similarly, the programme does not generate a differential impact in sustainability-related domains. In some areas, such as climate adaptation or access to technology for vulnerable groups, the effects are neutral or even negative. This is not a

single, well-defined problem, but rather a combination of interrelated challenges affecting both firms and evaluators, as well as the measurement framework itself.

First, firms applying to the programme have limited incentives to prioritise social impact. Given its historically low weight in the evaluation criteria, companies tend to focus their efforts on technological development and business viability, which are more decisive for funding decisions. This reduces both the quality and the comparability of the information provided on social impact.

Second, evaluators may face constraints in assessing this dimension. While CDTI's technical staff have strong expertise in scientific and technological domains, social impact requires different types of knowledge, often spanning areas such as inclusion, sustainability, or societal needs. This creates a gap between the complexity of the concept and the tools and expertise available to assess it consistently.

Third, and more fundamentally, social impact is difficult to define and measure. It is a multidimensional and inherently uncertain concept, which is not easily captured through standardised indicators. Existing measures rely largely on self-reported data and heterogeneous interpretations, introducing noise and limiting their reliability as a basis for decision-making.

These and other challenges are reflected in the programme's observed outcomes and suggest that the current assessment framework is not effectively identifying or prioritising projects with higher social impact potential. Taken together, this creates a structural gap between the growing policy ambition to promote socially impactful innovation and the ability of existing assessment systems to identify it. Addressing this gap requires not only better data or tools but also new approaches that can support more consistent, transparent, and evidence-based decision-making.

1.3 A new way of approaching problems

A shift in leadership at CDTI in 2024 marked a turning point in how the agency approached the challenge of improving the social impact of its programmes. This change brought a greater openness to experimentation and evidence-based approaches, creating the conditions for exploring new ways of strengthening the evaluation process.

In this context, CDTI became a Member of IGL, joining a broader international network focused on experimentation and the use of evidence in innovation policy. This marked the starting point of a new collaboration, not built around a predefined solution, but

around a shared objective: to explore how the assessment of social impact in Neotec could be improved in practice.

The collaboration developed progressively, beginning with a phase of joint exploration and ideation. Both organisations worked together to identify feasible entry points for experimentation within the institutional constraints of the programme. This included discussing alternative evaluation approaches, reviewing existing data sources, and exploring potential funding opportunities to support a more ambitious experimental design.

Alongside interest in experimentation, the collaboration was also driven by a broader question increasingly faced by innovation agencies: whether AI can meaningfully improve the design and delivery of policy programmes. In the context of the difficulties associated with assessing social impact, the focus was on evaluating AI's practical value as a support tool within a real administrative process, exploring whether integrating AI into the assessment of grant applications could improve consistency, reduce noise, or help evaluators identify relevant signals more effectively.

At the same time, there was interest in a more practical question: not only whether AI could help address some of these evaluation challenges, but also whether evaluators would actually use and respond to it in practice. More broadly, CDTI made a bold step attempting to move beyond simply digitising existing processes and towards identifying concrete use cases where AI can generate measurable improvements in core policy objectives.

2. Conditions to build the experiment

2.1 First steps

The initial ambition behind the collaboration was to explore whether more rigorous experimental methods could help improve the social impact of the Neotec programme. Early discussions focused on the possibility of implementing a large-scale Randomised Controlled Trial (RCT) linked to the programme's funding process.

However, as the project evolved, it became clear that directly testing the long-term social impact of grants through an RCT would face substantial practical and institutional barriers. The core issue was the use of randomisation in a public funding context. Legal concerns were raised around potential conflicts with principles of fair competition, particularly if funding decisions were influenced by experimental variation. There were

also concerns about potential legal challenges from applicants, as well as risks related to delays or disruptions in the evaluation process.

The approach also raised broader questions around feasibility. Measuring the downstream social impact of funded firms would require long time horizons, substantial resources, and complex political and administrative coordination. But rather than abandoning the initiative, the collaboration adapted its focus and experimental approach. Instead of attempting to evaluate the long-term social impact of grants directly, the project moved towards a more immediate and actionable question: whether alternative evaluation approaches, including the use of AI-assisted assessment tools, could improve the consistency and quality of project selection decisions.

This reframing proved critical. The focus moved towards identifying an alternative experimental design capable of preserving the core learning objectives while adapting to institutional, resource, and time constraints.

2.2 Adapting the design

Considering CDTI's institutional framework, the original plan was redesigned around the principle that if experimentation could not be integrated directly into funding decisions, it could still be conducted alongside them.

The solution proposed by CDTI's leadership was the development of a "shadow experiment", an evaluation exercise running alongside the standard Neotec selection process, but without influencing funding decisions. This design ensured that all applications would continue to be assessed under the existing rules, while alternative evaluation methods could be tested independently. Importantly, the shadow experiment did not run simultaneously with the official evaluation process. The experimental assessments only began once the formal evaluation of the 2025 applications had already concluded, ensuring that the exercise could not affect funding decisions in any way.

This transition also reflected a broader principle within evidence-based policymaking: experimentation can take different forms depending on the level of risk, realism, institutional constraints, and learning objectives involved. Alternative experimental approaches can still generate valuable operational learning in contexts where direct intervention in policy delivery is not feasible. In this case, the shadow experiment functioned as a controlled testing environment for alternative evaluation approaches within an existing public programme.

A staged approach to evidence-based policy

Five experimental methods, ordered from **highest control** (lab-like, low risk) to **highest realism** (real-world, scaled impact).

Stage & tool	Question it answers	Best for	Trade-offs
1 **Concept & theory** LAB EXPERIMENT	Why should this idea work in principle? What behaviour are we targeting?	Testing **behavioural mechanisms** in isolation, before designing a full intervention.	CONTROL **Very high** REALISM **Low** RISK **Very low**
2 **Design & optimisation** SHADOW EXPERIMENT	How can we best design and deliver this intervention?	Stress-testing **processes and operations** in parallel with live services.	CONTROL **Medium** REALISM **High (process)** RISK **Very low**
2 **Design & optimisation** RAPID-FIRE TRIAL	Which of several variants performs best?	Quickly iterating on **narrow design choices** (e.g. wording, layout, click-through).	CONTROL **Variable** REALISM **High (narrow)** RISK **Low**
3 **Pre-launch validation** EFFICACY TRIAL	Does the intervention show promise under ideal conditions?	Building **proof of concept** before committing to a full rollout.	CONTROL **High** REALISM **Medium to high** RISK **Medium**
4 **Real-world assessment** IMPACT EVALUATION (RCT)	What is the impact of the policy in practice, at scale?	Producing **robust evidence** to inform a final scale-up or stop decision.	CONTROL **Low to medium** REALISM **Very high** RISK **Potentially high**

Under this design, different methods for assessing social impact were applied to the same set of proposals. Their outputs could then be compared ex post to analyse consistency, identify potential biases, and evaluate which approaches were more aligned with observed outcomes. Crucially, because the shadow evaluations did not affect allocation decisions, the design avoided legal concerns related to fairness and competition.

This shift required a reframing of the experiment's ambition. While the original RCT aimed to test causal effects directly on funding decisions, the shadow experiment focused instead on understanding the behaviour and performance of different evaluation approaches. In this case, adapting the design was not a compromise that weakened the project, but a necessary step to make it possible.

2.3 The architecture to make it happen

The full development of the shadow experiment was the result of a carefully constructed collaboration between different actors with complementary roles. At its core, the process relied on a three-way interaction between a leading policymaker from CDTI, an academic lead from Humboldt University, responsible for the methodological design, and an IGL member acting as a knowledge broker.

Each actor brought a distinct perspective and set of capabilities. CDTI contributed institutional knowledge, facilitated implementation within the organisation, and provided access to programme data and operational processes. The academic partner developed the methodological approach, ensured research rigour and adherence to protocols, and led the analysis of results. IGL played a central role in connecting both worlds, coordinating the different actors involved, and driving the process forward throughout the different stages of the experiment.

A key function of IGL's role was the translation between academic and policy logics: adapting experimental concepts to fit within the operational realities of a public funding programme, while preserving enough rigour to generate meaningful evidence. It was also responsible for structuring and managing the project, ensuring alignment across stakeholders, and maintaining momentum throughout a process that evolved significantly over time.

In parallel, IGL played an active role in shaping the experimental design. This included refining the transition to a shadow experiment, developing an AI tool, and supporting the development of the protocols required for implementation. The management of external stakeholders, particularly the recruitment and coordination of social impact experts, was also a central component of this work.

But the process also required continuous mediation and conflict resolution. Experimentation within public agencies often requires navigating a range of institutional concerns that go well beyond methodological design. In practice, introducing new approaches such as AI-assisted assessments or external evaluators can raise questions around data protection, confidentiality, operational disruption, and the appropriate use of information. The experimental process in itself may also generate internal scepticism, particularly when it challenges established procedures. Addressing these concerns required continuous dialogue, adaptation, and trust-building across the different actors involved.

3. The experiment

3.1 Developing an AI tool

As part of the experiment, IGL developed a dedicated AI system to support the assessment of social impact in Neotec applications. The tool was designed using historical data from 283 approved projects between 2021 and 2023, combining proposal texts with information from end-of-project surveys to estimate the likely social impact of new applications.

The final system followed a two-step approach. First, a predictive language model analysed the executive summary and social impact sections of each proposal to generate a numerical social impact score on a 0–10 scale. Second, a generative AI model based on Google Gemini used both the proposal text and the predicted score to produce a written justification identifying relevant evidence from the application, while protecting data confidentiality.

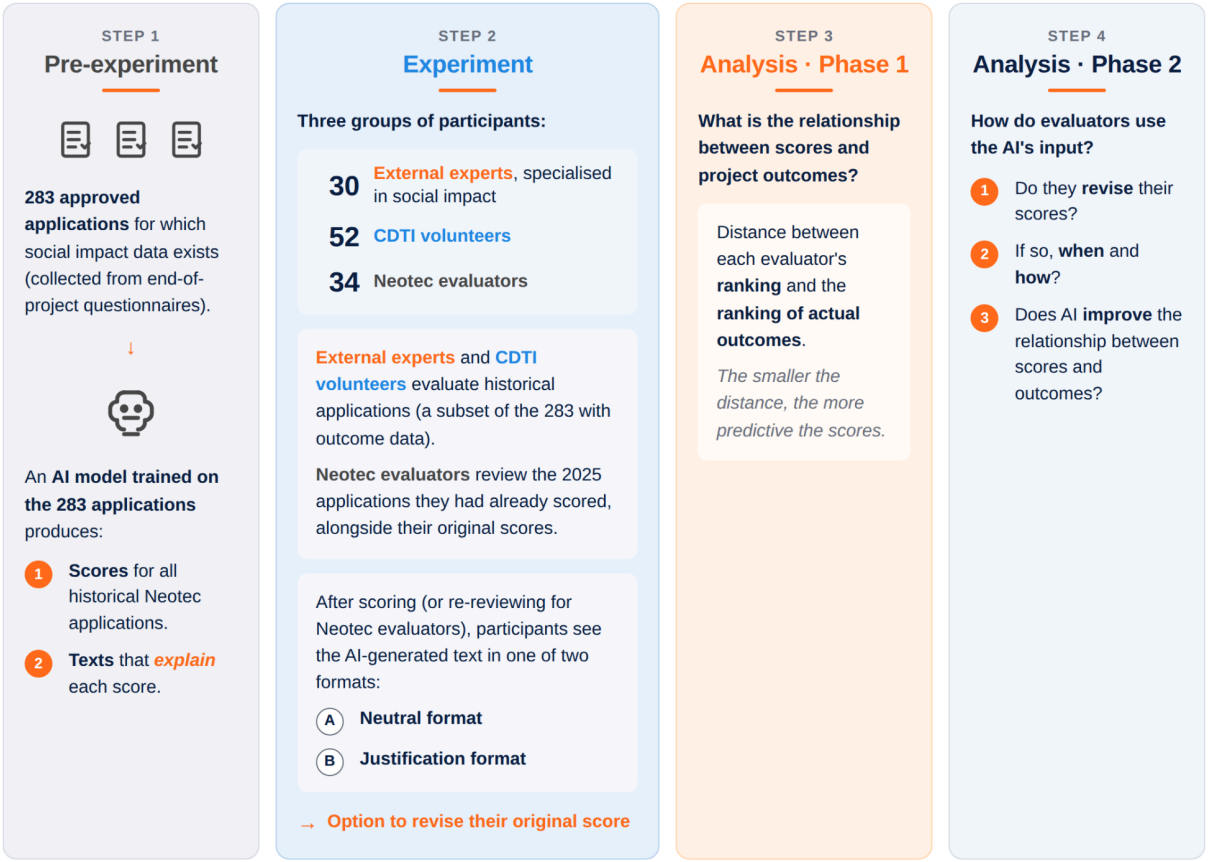
This sequential structure emerged after early development attempts showed that asking a single model to simultaneously generate both scores and explanations often produced unstable or weakly grounded outputs. To address confidentiality concerns, proposals were anonymised before being processed by the generative AI system. The experiment also tested two different styles of AI-generated explanations, allowing the analysis to explore how presentation formats influence evaluator behaviour.

3.2 Experimental design

The final design of the Neotec shadow experiment combined a structured evaluation exercise based on both historical proposals (from the 2021–2023 calls) and more recent applications from the 2025 call, allowing for the analysis of both predictive performance and behavioural responses.

The experiment brought together three distinct groups of evaluators to capture different perspectives on social impact assessment. These included CDTI's technical staff responsible for programme evaluation, a group of internal CDTI volunteers without direct experience in Neotec assessments, and external experts with specialised knowledge in social impact. This composition allowed the experiment to compare how professional evaluators, non-specialist insiders, and subject-matter experts approach the same task, providing a richer understanding of variation in judgement and potential sources of bias in the evaluation process.

How the experiment works



The experiment was organised into two main phases:

In **Phase 1 - Parallel evaluations**, proposals were evaluated independently by the three groups. The objective was to compare how different types of evaluators assessed social impact and to measure their predictive capacity, defined as the extent to which their rankings aligned with observed project outcomes derived from end-of-project surveys.[1] This phase enabled a direct comparison between internal technical assessments, external expert evaluations and AI-generated scores without any interaction between them.

[1] While these outcome indicators provide only a partial proxy for broader social impact, they constitute the most systematic source of post-project information currently available within the programme.

In **Phase 2 - AI feedback and score revision**, evaluators were exposed to AI-generated information and given the opportunity to revise their initial scores. This phase focused on understanding whether evaluators update their judgments when presented with new evidence, under what conditions they do so, and how the format of the information influences their decisions. A small share of proposals (around 5%) was randomly assigned to a pure control group without AI input, allowing for comparison.

3.3 Implementation

A key challenge during implementation was how to ensure a sufficiently large and diverse sample of evaluators. Early on, there were concerns that relying only on Neotec's technical evaluators would limit both the scale of the exercise and the variation needed for meaningful analysis. To address this, the pool of participants was expanded to include CDTI staff without prior experience in Neotec evaluations, allowing for a broader range of perspectives and increasing the overall sample. A registration process was opened for anyone within CDTI to join, and it was promoted internally through their communication channels.

In parallel, external social impact experts were incorporated into the experiment. These were recruited through a specialised consultancy, which supported the identification and onboarding of profiles with relevant expertise. At the same time, participation from Neotec evaluators was secured through direct, one-by-one engagement, reflecting both the importance of their role in the experiment and the need to ensure commitment in a context of high workload.

The final sample included 116 participants across three groups, representing 83% of all registered participants. The sample comprised 30 external social impact experts, 52 CDTI internal volunteers not usually involved in Neotec evaluations, and 34 technical evaluators directly linked to the programme. Together, participants reviewed 348 proposals, producing a total of 828 individual evaluations.

To support the exercise, the academic lead developed a dedicated digital platform tailored to the needs of the experiment. This platform integrated the evaluation workflow, allowing participants to access anonymised proposals, assign scores, and review AI-generated outputs within a single interface. It also enabled the random assignment of treatments, controlled the exposure to different AI formats, and ensured that all interactions were recorded for subsequent analysis, while protecting data privacy. From an operational perspective, the platform was a critical component, as it allowed the

experiment to be conducted in a structured and standardised way across all participant groups.

Despite these preparations, there were persistent concerns about participation rates and the risk of not reaching the desired sample size. To mitigate this, as well as to build awareness and support across the organisation, IGL and CDTI organised an in-person evaluation session, bringing together participants in a single setting. During this session, evaluators worked simultaneously for approximately two hours at CDTI's premises, completing a significant share of the assessments. This collective format helped ensure sufficient data collection, with around 70% of participating CDTI staff completing their evaluations during the session, while also facilitating coordination and real-time support.

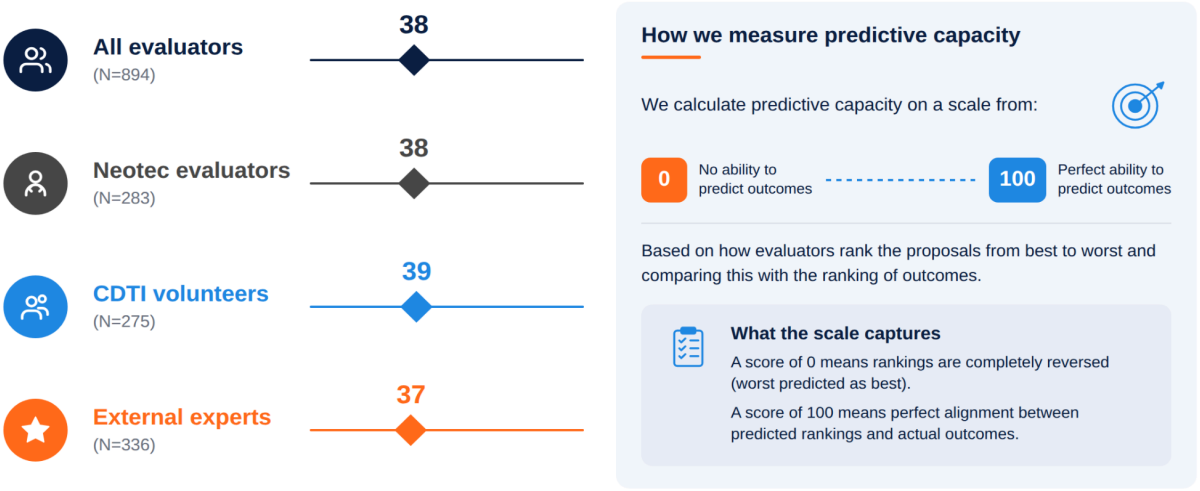
Session at CDTI's Function Room in Madrid,18/12/2025.

The implementation also raised several concerns related to data protection and confidentiality. These included questions around the level of anonymisation of proposals and the use of AI systems in handling potentially sensitive information. Addressing these issues required additional safeguards in the design of the platform and the handling of data. While these concerns introduced complexity, they were ultimately resolved in a way that allowed the experiment to proceed without compromising institutional or legal requirements.

3.4 Main results

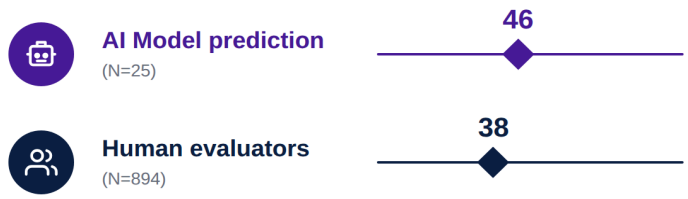
The experiment provides a set of consistent findings across three dimensions: the predictive capacity of current evaluation systems, the contribution of AI, and how evaluators respond to new information.

All evaluators have limited predictive capacity



A first and striking result is that predictive capacity remained limited across all evaluator groups when assessing social impact. Evaluators showed difficulties consistently ranking projects in line with observed outcomes, and this pattern was relatively similar across technical evaluators, external experts, and internal volunteers. This suggests that the challenge is not primarily driven by familiarity with the programme or specialised expertise in social impact assessment, but reflects a broader difficulty in anticipating the future societal effects of early-stage innovation projects. Essentially, these findings highlight the high level of uncertainty involved in assessing social impact, particularly when relying on partial and imperfect outcome indicators.

The AI model may have greater predictive capacity



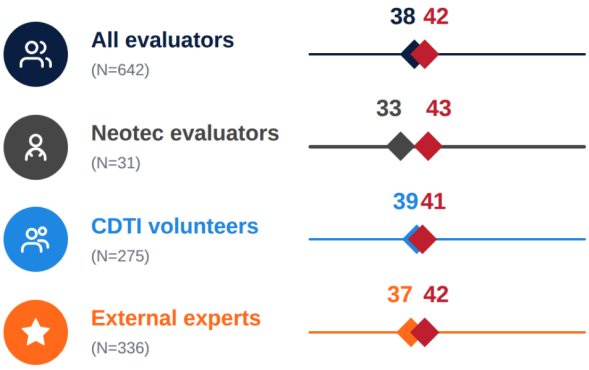
What this comparison shows

By calculating the predictive capacity of the scores produced by the AI model, we obtain a slightly higher metric than that of human evaluators.

Limitation: this is based on a separate and very small sample (N=25).

The AI system shows a somewhat higher predictive capacity than human evaluators, although the difference remains modest and should be interpreted with caution due to the smaller sample used for validation. The results suggest that AI can identify patterns in the available data that are not consistently captured by human evaluators. While the experiment relies on imperfect proxy measures of social impact based on end-of-project survey data, these findings indicate that AI may provide useful complementary signals to support decision-making under conditions of high uncertainty.

AI can improve evaluators' scores



What this comparison shows

Without AI With AI assistance

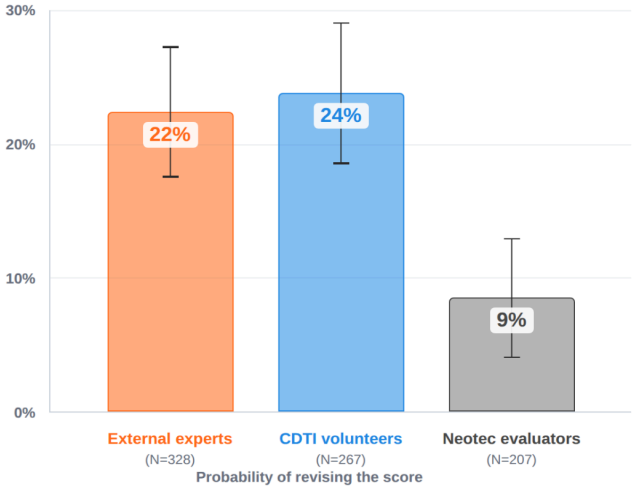
Across every evaluator group, predictive capacity increases when scores are produced with AI assistance, with the largest gain observed among Neotec evaluators.

Limitation: for Neotec evaluators, we only have a sample of 31 proposals from the 2021–2023 period. They primarily reviewed their 2025 proposals, for which we do not yet have outcome data.

Exposure to AI-generated information leads to small but consistent improvements in evaluators' predictive capacity. Across all groups, the average distance between predicted and observed outcomes increases after exposure to AI-generated assessments. However, behavioural responses remain limited overall: in around 80% of cases, evaluators do not revise their initial scores after receiving AI input. Taken

together, this evidence suggests that the improvements associated with AI use are directionally consistent but modest in magnitude.

Evaluators use AI selectively



When evaluators revise their scores

Evaluators are more likely to revise their scores when the AI:

a. Suggests **lowering the score** (fear of false positives).

b. Suggests a score **further from their own** (novel and useful information).

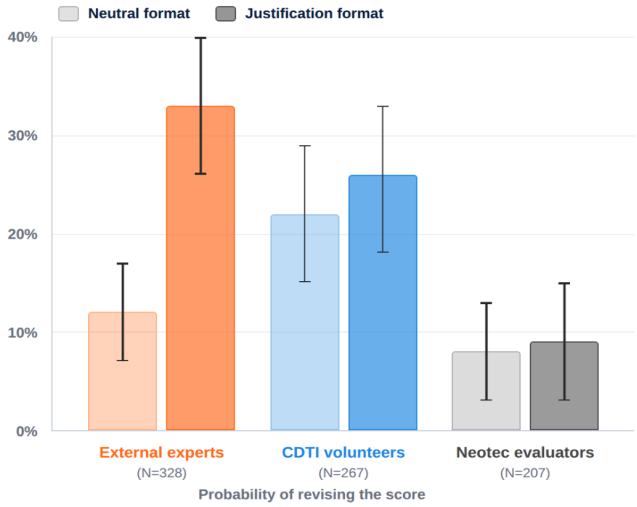
c. Provides a **correct suggestion** (closer to actual outcomes).

Responses to AI varied substantially across groups. External experts and CDTI volunteers revised their initial scores significantly more than Neotec evaluators. The difference was even clearer at evaluator level: 75% of Neotec evaluators kept all their original scores unchanged, compared with 20% of external experts and 35% of CDTI volunteers. This suggests that evaluators more embedded in the programme may rely more strongly on their prior judgement and established evaluation routines, while others are more receptive to external input.

The likelihood of revision also depended on the nature of the AI signal. Evaluators were more likely to revise when the AI suggested a score further from their own: they changed around 25% of scores when the disagreement was relatively large, compared with 14% when the two scores were closer. This suggests that AI input was more influential when it offered potentially novel and useful information that differed substantially from the evaluator's initial judgement, rather than merely pointing towards a marginal adjustment. They were also more likely to revise when the AI provided a more accurate suggestion, one closer to the observed project outcomes than their own initial assessment, with revision rates of 25%, compared with 16% when it did not. Finally, evaluators were more likely to reconsider their judgement when the AI suggested lowering rather than raising

the score. In essence, these patterns suggest selective rather than automatic use of AI input.

Design matters: the AI format changes behaviour



Two AI presentation formats

We present the AI-generated text in one of two formats:

1. **Neutral format** summarises the evidence identified by the AI (facts only).
2. **Justification format** summarises the same evidence but uses persuasive language (value judgements).

A key result is that the format of AI-generated information can significantly affect evaluator behaviour. We randomly assigned proposals to one of two presentation formats and found that more explicit and persuasive explanations increased the proportion of scores revised from 14% to 24%, compared with a neutral format. But this overall effect was driven mainly by external experts. This shows that the influence of AI also depends on how its outputs are communicated to users.

The interaction between format and the direction of the AI signal revealed a further asymmetry. When the AI conveyed a less favourable assessment, downward adjustments occurred under both formats. However, when the AI conveyed a more favourable assessment, the average change under the neutral format was close to zero, while more persuasive explanations led to more frequent and larger upward revisions. Essentially, neutral evidence appeared sufficient to prompt downward adjustments, whereas upward revisions required stronger justification. This points to a possible conservative tendency: depending on how its outputs are presented, AI support could make a funding process more conservative even without being explicitly instructed to do so.

4. Lessons

4.1 Lessons from the results

Predicting social impact is inherently difficult: A first lesson from the experiment is that predicting the future social impact of innovation projects is inherently challenging. All evaluator groups showed low predictive capacity, regardless of their background or level of expertise.

External experts specialised in social impact did not substantially outperform other evaluators, suggesting that the main difficulty lies not only in evaluator knowledge, but in the broader challenge of defining, measuring, and anticipating social impact outcomes in a consistent way. More broadly, these findings point to a structural measurement problem that extends beyond CDTI and reflects wider international debates on how to assess societal impact within innovation policy.

AI works as cognitive support, not as an oracle: AI-generated information produced small but consistent improvements in predictive performance, but evaluators used it selectively rather than following it automatically. They were more likely to revise when the AI offered a score further from their own and, in historical cases, when its suggestion was closer to observed outcomes. They were also more receptive to signals pointing towards a lower score, although experienced Neotec evaluators were particularly unlikely to revise.

Presentation mattered too. Explicit and persuasive explanations increased the proportion of scores revised. Combined with the greater receptiveness to downward recommendations, this means that interface design could unintentionally make funding decisions more conservative. The main value of AI therefore lies in providing an additional signal that prompts evaluators to reconsider their judgement, not in replacing it. Realising that value also requires institutional conditions that make revising an initial assessment legitimate.

This brief presents the main high-level lessons. The full technical report prepared by CDTI contains the broader analysis of behaviour, AI performance, scoring dynamics, treatment effects and methodological robustness. The academic team is also preparing a research paper with a more complete account of the methodology and results.

4.2 Lessons from the overall experiment

Building experimentation capacity requires more than technical expertise: One of the clearest lessons from the project is that implementing experimentation in public administration is not only a methodological challenge, but also an organisational and relational one. Many of the institutional concerns raised throughout the process, around data protection, fairness, operational risks, or the use of AI, were legitimate and needed to be addressed carefully. However, managing these concerns required sustained coordination and negotiation work that extended well beyond the technical design of the experiment itself.

This highlighted the importance of having a dedicated intermediary role capable of connecting academic, operational, and institutional perspectives. As CDTI pointed out, *"the collaboration with IGL was critical in developing strategies to navigate the practical and institutional challenges involved in running an experiment of this kind, while also supporting the connection between the academic and operational sides of the project."* Separating these functions from the academic lead also proved important in practice, allowing methodological work to remain focused while institutional challenges were managed in parallel.

Flexibility and trust are essential for successful experimentation: A second lesson is that experimentation in real policy settings rarely follows the original plan exactly. The project required more time, adaptation, and coordination than initially expected, and several aspects of the final design differed substantially from the initial proposal. Rather than undermining the process, this flexibility became one of the main reasons the experiment succeeded.

The ability to adapt depended heavily on trust and engagement across all actors involved. The academic lead was entrusted with substantial responsibilities, including data handling, analysis, platform development, and AI system design, requiring dedicated and continuous work. At the same time, high-level engagement from within the policymaking institution proved critical. Internal champions within CDTI actively defended the project, encouraged participation, and helped navigate internal resistance and operational concerns.

The project also illustrates the importance of pragmatic experimentation. Instead of abandoning the initiative when initial plans became unfeasible, the collaboration searched for workable alternatives adapted to institutional realities. This included redesigning the experiment as a shadow exercise, developing a dedicated platform to address confidentiality concerns and control experimental variation, and organising an

in-person event to secure sufficient participation. These solutions were not part of the original design, but emerged through iterative problem-solving during implementation.

Overall, the experience shows how experimentation in government requires technical expertise, institutional entrepreneurship, coordination capacity, and a willingness to adapt designs to real-world constraints.