Evaluating AI homelessness interventions:

A practical guide for local government

Luke Arundel, Julia Ellingwood, Isobel Harrop and Michael Sanders

Centre for Homelessness Impact

October 2026

Contents

Key recommendations at a glance 5

Is this guide for me, and what will it help me decide? 7

What does this guide cover, and what does it not? 12

Why should I test AI rather than just implement? 15

What problem is the AI tool solving, and is AI the right answer? 17

What would count as evidence this tool is working? 24

How do I run a fair test of the AI tool? 30

What if running a randomised controlled trial is not feasible or proportionate? 32

Is the tool being implemented as intended? 38

Is the tool worth what it costs, and what costs and benefits could I be missing? 42

What could go wrong, and what do I need in place before I commission? 45

How do I share what I've learned, and learn from what others are doing? 48

Closing recommendations 50

References 52

Endnotes 54

Appendix A: Theory of Change 55

Appendix B: Trial design 56

Appendix C: Safeguarding flow 61

Purpose

The purpose of this guide is to help local authorities make evidence-informed decisions about commissioning, evaluating, scaling or stopping Artificial Intelligence (AI) tools in homelessness prevention. It provides practical guidance on how to evaluate these interventions in a way that is proportionate, robust and sensitive to the specific risks of working with people at risk of homelessness. The guide encourages local authorities to move beyond short-term metrics, such as usage or satisfaction, and focus on meaningful outcomes, including homelessness duties, sustained tenancies, safety, equity, cost-effectiveness and real-world implementation.

This is a practical guide, not an evaluation report, and does not present findings. It sets out what to do, and what matters, when evaluating AI tools you use or commission, using our ongoing trial of one AI housing advice tool as a worked example throughout. Findings from that trial will be published separately.

About the Centre for Homelessness Impact

The Centre for Homelessness Impact (CHI) is an independent organisation that champions the creation and use of better evidence for a world without homelessness. Our mission is to improve the lives of those experiencing homelessness by ensuring that policy, practice and funding decisions are underpinned by reliable evidence.

About the Policy Institute

The Policy Institute at King's College London (KCL) works to solve society's challenges with evidence and expertise. We combine the rigour of academia with the agility of a consultancy and the connectedness of a think tank. Our research draws on many disciplines and methods, making use of the skills, expertise and resources of not only the institute, but the university and its wider network too.

This work was funded by the UK Government Evaluation Accelerator Fund (Phase 4, 2025-26).

Key recommendations at a glance

1. **Before you implement, ask whether AI is the answer.** Carefully consider the value an AI tool adds on top of already available housing and homelessness tools and services, such as existing guidance, phone lines or webpages.

   → See 'What problem is the AI tool solving, and is AI the right answer?'

---

2. **Build in evaluation and aim for the strongest design you can support.** The digital nature of these tools makes randomisation easier than you might expect.

   → See 'How do I run a fair test of the AI tool?'

---

3. **Measure outcomes that matter, and expect a lag** - homelessness duties, sustained tenancies, contacts with the council, and whether outcomes for users improved, are as important as not just user satisfaction and hours saved. Plan and budget for longer-term follow-up.

   → See 'What would count as evidence this tool is working?'

---

4. **Put the essentials in place, whatever evaluation design you use** - a Theory of Change, assurance that the tool is safe and performs the core tasks it is designed for, and minimum monitoring data: usage and reach, safety incidents and escalations, and ongoing accuracy checks.

   → See 'What if running a randomised controlled trial is not feasible or proportionate?'

5. **Count the full costs and benefits** - including staff time, integration, and ongoing upkeep. Include benefits to both the council, such as what a prevented homelessness duty saves in temporary accommodation spending, and to users, such as improved housing stability and wellbeing.

   → See 'Is the tool worth what it costs, and what costs and benefits could I be missing?'

---

6. **Expect the tool to change mid-evaluation.** Keep a meticulous change log and ongoing accuracy monitoring of the AI tool. This is particularly relevant to the homelessness sector, where statutory obligations and best practice will likely change over time.

   → See 'Is the tool being implemented as intended?'

---

7. **Treat chatbot logs as sensitive data.** If using generative AI, require human-in-the-loop escalation for high-risk cases, complete a Data Protection Impact Assessment, and consider an Algorithmic Transparency Recording Standard record.

   → See 'What could go wrong, and what do I need in place before I commission?'

---

8. **Share what you find,** including null and negative results, and collaborate with other councils testing similar AI tools.

   → See 'How do I share what I've learned, and learn from what others are doing?'

Is this guide for me, and what will it help me decide?

Introduction to the guide

Local authorities are adopting artificial intelligence (AI) tools at pace. In 2025, the Local Government Association (LGA) published the second instalment in their **State of the Sector: Artificial Intelligence** report series, where they reported that among the councils surveyed, 95% are using or exploring AI tools, with generative AI being the most widely adopted type (Local Government Association, 2025b). Applications of AI tools in housing and homelessness already range from supporting **temporary accommodation allocations**, to taking **case notes**, to **predicting homelessness**.

There is also a push for the adoption of AI from central government, with the Ministry of Housing, Communities and Local Government (MHCLG) recently introducing their Local AI team, working with councils to support the responsible use of AI across local government. In the context of budgetary constraints (particularly with regard to prevention), and increasing statutory requirements, there is a strong case for understanding how AI can be deployed effectively in local authorities to prevent homelessness.

However, adoption is outpacing evidence. There is not currently enough evidence on the effectiveness of AI tools (Hauser et al., 2025; Parker et al., 2025). Local authorities have identified this problem, with 58% reporting that lack of evaluation of AI tools presented a barrier to procurement (Local Government Association, 2025b). This is an issue because AI is not cost- and risk-free to use. Setting aside the immediate costs of licensing, hardware and software requirements, AI can incur additional costs for local authorities in the form time spent by staff learning and implementing new solutions, dealing with errors stemming from hallucinations or poor administration of tools, and potential loss of trust among staff or service users resulting from unsuitable tools or poor deployment.

In the face of these costs and risks, it is essential that local authorities have reassurance that the AI tool being considered is a sound investment, is safe to use, and delivers the outcomes we care about. An important piece of this will be commissioning more independent evaluations of AI effectiveness. The Evaluation Task Force and HM Treasury have published guidance on running impact evaluations of AI interventions, as a new annex to their Magenta Book guidance (Evaluation Task Force & HM Treasury, 2026). Their guidance is designed to be broadly applicable, spanning many different types of AI technologies, implementation contexts, and outcome measures.It provides key principles of best practice that can apply generally.

Here we aim to help bridge the gap between the broader guidance on AI and evaluation, and the use of AI in the homelessness sector. We do this by focusing on practical recommendations and insights, and the unique methodological, ethical and operational considerations for evaluating AI in the homelessness sector specifically.

Table 1. Navigating this guidance

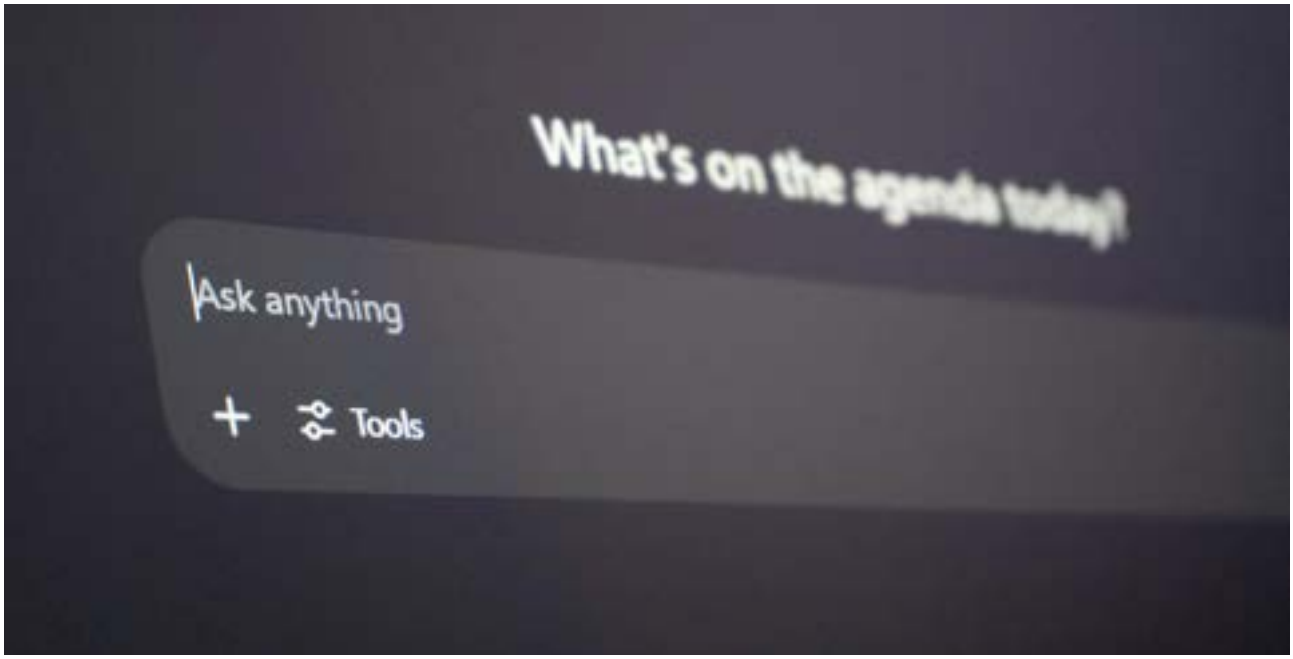
If you are…	Go to…
Deciding whether AI is the right answer	What problem is the AI tool solving, and is AI the right answer?
Deciding what success would look like	What would count as evidence this tool is working?
Designing the evaluation	How do I run a fair test? and What if an RCT is not feasible or proportionate?
Understanding how the tool is used in practice	Is the tool being implemented as intended?
Assessing value for money	Is the tool worth what it costs?
Considering safety, data and governance	What could go wrong, and what do I need in place before I commission?
Sharing what you learn	How do I share what I've learned?

We set out the steps to be taken when planning an evaluation of an AI tool, drawing on the case study of Housing Helper (see Box 1), an AI chatbot intended to give housing advice to people at risk of homelessness. We emphasise the importance of establishing a clear Theory of Change, selecting outcome measures that reflect meaningful changes in users' circumstances, and employing evaluation designs capable of generating robust evidence of effectiveness, giving particular attention to issues that are relevant to AI-enabled interventions.

We recommend randomised controlled trials (RCTs) as the preferred approach for estimating impact, supported by implementation and process evaluations to understand how interventions are delivered and why they succeed or fail. Where experimental approaches are not practicable or proportionate, or where your authority does not have capacity to support them, we outline alternative evaluation designs. We also set out the minimum that every local authority should have in place for any AI homelessness tool it deploys, whatever the evaluation design, to ensure the tool is safe, accurate, and reaches the people it was commissioned to help.

We consider how much evidence is enough for a given decision - not every decision needs the same evidence, or the same amount of time. Whether the tool is safe to keep running, whether it is being used, and whether it is reaching the people it was bought to help are decisions you can make within weeks or months. Whether to scale it across your service, or recommend it more widely, likely depends on whether it improves outcomes.In homelessness that will often take at least a year. Rather than choosing between speed and rigour, match the evidence to the decision you need to make.

The overarching principle is that decisions to adopt, scale or discontinue AI-enabled homelessness interventions should be informed by evidence. Robust evaluation is essential to ensure that these tools improve outcomes for people experiencing or at risk of homelessness, represent value for money, and operate safely, ethically and transparently within the public sector.



Our worked example - the Housing Helper evaluation

Housing Helper is an AI housing advice chatbot, developed by the social enterprise Change Please. It works by assessing a service user's housing situation, offering targeted housing advice, and connecting users to relevant resources and services. Housing Helper can also draft custom letters and emails, for example letters to a landlord asking for a temporary rent reduction or flexible payment plan in response to financial difficulties, or letters to bailiffs to appeal for consideration of mitigating circumstances. Housing Helper works in 55 languages and can translate letters and documents to make them accessible for people with English as an additional language.

Our evaluation, (funded by the Evaluation Task Force through the Evaluation Accelerator Fund) primarily seeks to understand whether Housing Helper reduces the likelihood that users are owed a homelessness duty. To do this, we have partnered with Southwark Council, who connected us to approximately 9,000 households deemed at risk of homelessness (e.g. by being in rental or council tax arrears), but are not currently in temporary accommodation. Because we cannot mandate use of Housing Helper, we have opted for an individually randomised encouragement design, where half of households in the sample are randomly selected to receive a set of "encouragement" messages (email, SMS and WhatsApp), inviting them to use Housing Helper (treatment), and the other half are sent an email from Southwark Council, directing them to existing information and resources (control).

What does this guide cover, and what does it not?

Scope and signposting to existing resources

This guide provides an overview of what to consider when evaluating AI tools in the homelessness sector specifically. Such tools broadly fall into three groups: 1) those used directly by residents, such as advice chatbots; 2) those used by staff, such as case-note summarisation or triage support; and 3) those that inform decisions about residents, such as predictive risk models. We focus on the first group, and on generative AI[1], using our Housing Helper evaluation as a case study, although many of the principles will also apply to these other tools.[2] The **AI Playbook** and the **Magenta Book AI Annex** already provide substantial coverage of AI more broadly, as well as the relationship between AI and evaluation. This guide should be treated as a companion guide to these other resources (see Table 2).

What is evaluation in this context?

'Evaluation' can mean two different things when applied to AI. Model evaluation – also known as assurance, or testing – asks whether a tool works as intended, looking at whether outputs are accurate, if it performs the task it is designed for, and whether it behaves safely. Impact evaluation focuses on whether deploying the tool changes outcomes for people. This guide focuses particularly on impact evaluation, although model evaluation is a prerequisite before moving forward to any kind of impact evaluation. If a tool is not accurate and stable, then it will be difficult to attribute any effect to it and, more importantly, an inaccurate tool should not be in front of people at risk of experiencing homelessness. The Magenta Book AI Annex takes a similar approach, excluding technical benchmarking from detailed discussion while noting that this is a vital activity for establishing that AI systems are effective and safe.

A developer should have done substantial testing before offering a tool to local authorities. Nonetheless, local authorities should always ask what kind of model evaluation has been done and make a judgement on whether it is sufficient for a homelessness context specifically. Further checks on real-world use will likely be needed before rolling it out more widely (see Table 2 for resources on model evaluation). Model evaluation is thus a necessary, but not sufficient, part of understanding the impact of an AI tool. Impact evaluation picks up where model evaluation ends, looking beyond whether the tool gives good answers, and towards whether giving people access to that tool causes a change in their outcomes.

Table 2. Where to go for AI resources

If you want…	Read…
What is specific to evaluating AI tools for people at risk of homelessness	This guide
General guidance on using AI in government	AI Playbook for the UK Government
Practical help buying AI as a council, including equality and data protection duties	LGA's Responsibly buying AI; ICO guidance on AI and data protection
How to test whether an AI tool is safe and accurate	DSIT's Introduction to AI assurance and Portfolio of AI assurance techniques
How to evaluate AI interventions more widely	Magenta Book AI Annex
Detailed statistical methods for trials	CHI's statistical guidance for evaluators; Designing and Delivering Randomised Trials for Social Policy
Running an implementation and process evaluation	CHI's IPE guidance
Economic evaluation methodology	The Green Book; CHI's economic evaluation guidance
Reflections on AI and the public sector more widely	Ada Lovelace Institute's Measuring Up and Learn Fast and Build Things

Why should I test AI rather than just implement it?

Why AI for homelessness needs particularly careful evaluation

There is growing pressure to adopt AI in local government. The government's **AI Opportunities Action Plan** commits the public sector to "push hard" on AI adoption, and discusses a "Scan, Pilot, Scale" approach - identifying promising uses, piloting them, and scaling what works (Department for Science, Innovation and Technology, 2025). As with other interventions, piloting and evaluating AI interventions is important because it ensures that the time, effort, and money the public sector puts into trying to make people's lives better actually does make people's lives better. In homelessness, for example, robust evaluation built the evidence base for Housing First, a model that provides housing immediately rather than requiring preconditions such as sobriety or employment. That evidence is one reason support has shifted towards Housing First in recent years (Teixeira, 2022).

AI interventions aimed at people at risk of, or experiencing, homelessness particularly need evaluating. This is because the risk of harm if an AI chatbot gives poor advice to someone in a housing crisis is far greater than, say, someone trying to renew their passport. The populations these tools are aimed at also differ from the population as a whole in important ways that matter for both design and evaluation. For example, people at risk of homelessness are disproportionately digitally excluded, which shapes who can use an AI tool at all, and consequently where the benefits and harms from the tool fall (Ofcom, 2025).

Some people experiencing homelessness also have specific safeguarding needs, and local authorities have statutory duties to prevent and relieve homelessness. This means that AI tools in homelessness will have to perform important functions beyond giving accurate and accessible information, including risk-triaging people presenting with very complex cases, and making accurate referrals to relevant local authorities or other services.

The high risk of doing harm, alongside the many complex tasks that are required to be useful and to keep people safe, means that evaluating the efficacy of AI tools in homelessness is ethically necessary, not optional, if these tools are going to be deployed and scaled.

In short, this is why we test before we scale. National policy strongly encourages AI adoption, but the evidence on effectiveness is limited, and homelessness is a domain where the harm from deploying AI badly can be severe.

What problem is the AI tool solving, and is AI the right answer?

Creating a Theory of Change

Before planning an evaluation, we should consider whether an AI tool is the best way to solve the problem at hand. Using AI simply for the sake of using AI should be avoided - it must be a good fit for the identified problem and not present unreasonable risks to users (Homeless Link, 2026). It is useful to think through what a non-AI version of the intervention - for example a leaflet, a phone line, or a webchat with a human - could not achieve that the proposed AI tool would. We will also revisit this idea in the **'Active controls' section in Appendix B.** If a well-written webpage would achieve the same thing as an AI tool, then the webpage would likely be the value-for-money option.

Creating a Theory of Change is helpful for thinking this through, and is the first step in planning any evaluation. It helps us to understand what the intervention aims to do, so we can select sensible outcome measures and design the evaluation appropriately. There should be a clear, logical pathway between use of the tool and the outcomes, and those outcomes need to be measurable and changeable within the given time period of the evaluation. If you can't articulate a credible Theory of Change for your AI tool, it is probably the wrong intervention for the problem you are trying to solve. Stop there, before deploying or evaluating it.

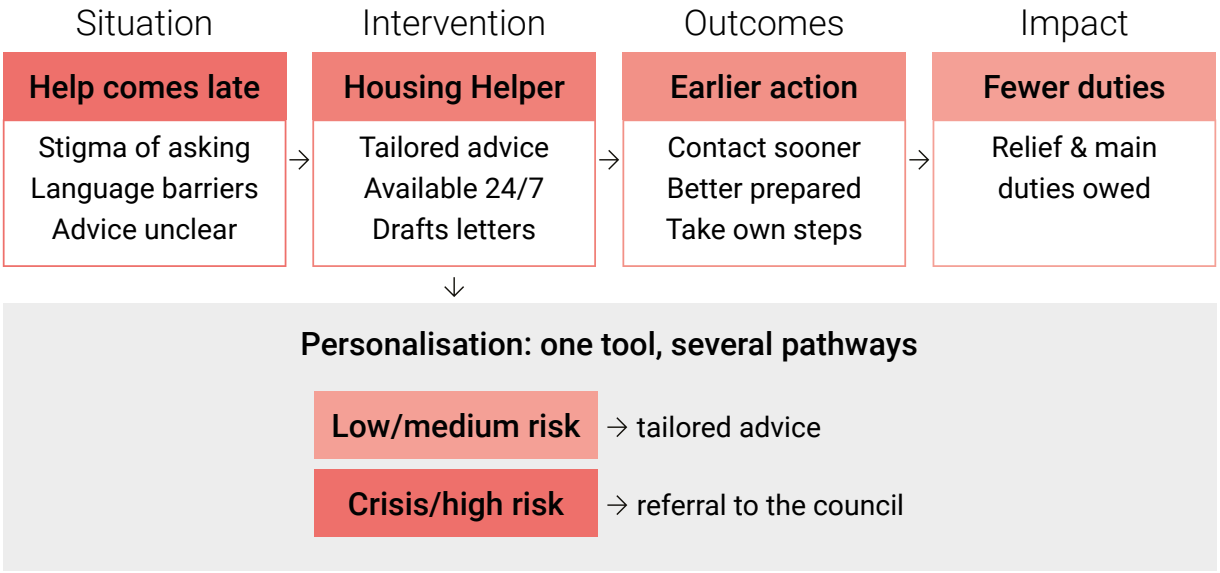
There are many resources available to help you to create a Theory of Change. We recommend Nesta's Theory of Change Toolkit. There are, however, particular considerations to keep in mind when creating a Theory of Change for a homelessness AI tool:

Who builds the Theory of Change

A Theory of Change built only with the tool's developer will reflect only the developer's assumptions. The Ada Lovelace Institute's review looking at claims about AI in the public sector notes that the dominant methodologies and data sources tend to reflect the perspectives of technologists, evaluators and AI companies, while the workers who use the tools, the frontline managers who implement them, and the people who experience the resulting service changes are frequently absent from research design (Deshmukh & Davies, 2026). For a homelessness tool, the views of these groups are likely to be particularly important. Housing officers will know what 'business-as-usual' looks like and whether a referral from a chatbot is usable. People with lived experience know whether the assumptions about trust, digital access, and willingness to disclose to a machine are realistic for people at risk of experiencing homelessness. The Theory of Change should be built with the vendor, housing officers, and people with lived experience of homelessness.

In the Housing Helper evaluation, the Theory of Change was developed jointly by the evaluation team, Change Please, Southwark Council, and CHI, and the tool was tested by CHI's lived experience associates before launch. Figure 1 presents a simplified version of the Theory of Change (for the full Theory of Change, see Appendix A).

Figure 1. Housing Helper's Theory of Change (simplified)



Personalisation

If one of the features of the AI tool you are evaluating is that it offers a personalised experience, then it is important to think through the following:

- The different pathways that different users may experience when using the tool, and whether users on different pathways would be expected to have different outcomes.
- How users will be triaged to different pathways, so we know what outcomes to expect for specific user types.
- What will count as having used the tool.

The third point is important because having users on multiple different pathways could make it harder to tell whether the tool is working. For instance, a very casual user who only plays around with a couple of prompts before logging out might not constitute "use" according to the Theory of Change. Including these types of users in the analysis could make it less likely that we can estimate an effect. Deciding on a threshold of what constitutes "using" the tool can help to mitigate this (e.g. features used, time engaging with the tool).

In practical terms, a high-risk household facing bailiff action and a low-risk household querying a tenancy clause are receiving essentially different interventions from the same tool, and your Theory of Change needs to account for both.

Model drift and adaptive systems

If the AI tool you are evaluating is designed to adapt to changes, such as new legislation or updated sources, it may become less accurate over time if those updates fail (we call this 'model drift'). It is important to consider how you will measure usefulness or accuracy over time for the evaluation. Drift also creates an attribution problem that is particular to AI evaluations in this sector: the outcomes that matter (homelessness duties) can take a year or more to materialise (see Time lag, below), so a tool that has changed substantially mid-evaluation complicates the question of which version of the tool produced the effect you observe. This is one reason the change log discussed below is so important (see 'Is the tool being implemented as intended?').

Case Study - Creating a Theory of Change for the Housing Helper evaluation

At the outset of the Housing Helper evaluation, we developed a Theory of Change in partnership with Change Please and Southwark Council (see Figure 1 and Appendix A).

We start with the articulation of the **situation** and the problem to be solved. The situation is that local authorities have a duty to offer housing advice and to support residents who are at risk of homelessness, and the problem is that the number of residents in need is increasing, putting pressure on local authorities and potentially delaying help until they reach a crisis point. In response to this situation, Housing Helper aims to offer targeted, early, accessible, and personalised housing advice at scale; identify high-risk cases and make triage referrals to the council; reduce homelessness duties owed overall; and reduce local authority time and money spent on crisis response.

After identifying the **situation/problem** and the aims of the tool, we worked backwards from there to identify the specific **outcomes and impact** of using the tool, then the **inputs and activities** required to realise those outcomes and impact. This "starting from the end" approach ensures that the aims remain the focus, and the logic path is built to achieve those aims. Finally, we reflected on and documented the underlying **assumptions** of how Housing Helper is theorised to work.

Considering the different pathways that users could take when using Housing Helper was a key part of this process: users might have accessed Housing Helper to get advice on a housing issue, to interpret or draft a letter, or to calculate the benefits they are owed. They might be offered a referral to Southwark Council, or they might be signposted to a charity. Thinking through these pathways enabled us to think about how we would measure whether these actions were taken, how successful those actions were, and what the impact of the actions was.

The large number of ways that Housing Helper can be used makes it more difficult to estimate the effect of the intervention. We decided to create a broad definition of what using Housing Helper meant for this evaluation, so that we could include as many users and ways of using Housing Helper as possible in our intervention:

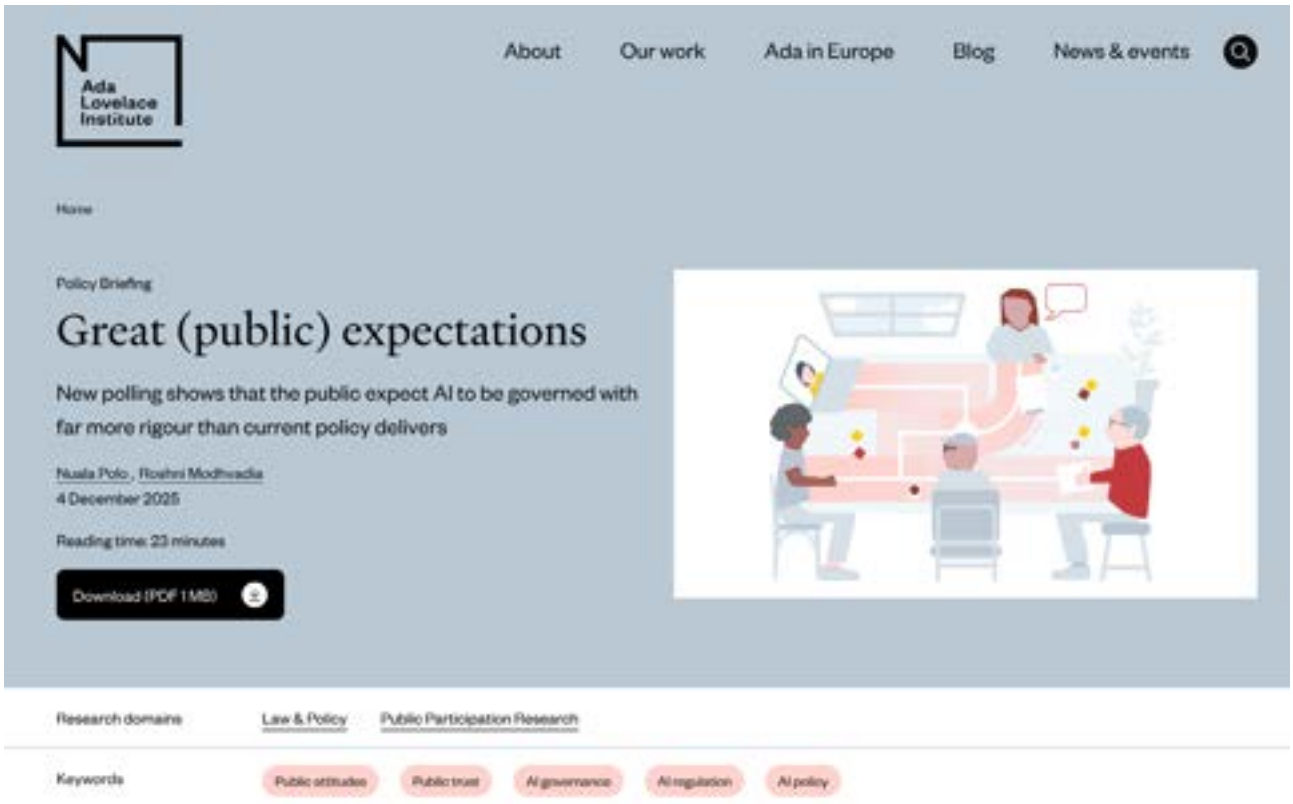
- Clicking the link provided in an encouragement message
- Completing the user sign-up flow and creating a Housing Helper account
- Using Housing Helper for a minimum of four minutes

Four minutes was suggested by Change Please as the minimum time required to be risk-triaged by Housing Helper and for the user to derive value of some kind. This is based on their experience using Housing Helper in the field (outside the context of this trial and Southwark borough).

We also reflected on whether and how Housing Helper would change and adapt over time and with use. Housing Helper relies on a retrieval-augmented generation (RAG) system that sits above a large language model (LLM), drawing upon publicly accessible, authoritative, and ethically approved sources like Shelter, Advice Now, and GOV.UK. A core assumption underlying the effectiveness of Housing Helper is that it is providing advice that is accurate and trustworthy in the first instance (i.e. it will not hallucinate), and in the second, that it will update its advice in response to new legislation, statutory requirements, and as sources change (e.g. contact information for signposted organisations). Failure to adapt to these changes would mean that the tool would get less useful over time, which would then affect outcomes and impact. This reflection exercise highlighted the importance of measuring the accuracy of Housing Helper throughout the course of the trial, rather than only in its development phase.

Trust

One key aspect to consider when developing a Theory of Change for AI tools in the homelessness sector will be trust: do the people using the tool trust its outputs?[3] If people do not trust a tool they will use it less. A report from the Ada Lovelace Institute on public attitudes to AI notes that, "the public feel disenfranchised and excluded from AI decision-making, and mistrust key institutions" (Polo & Modhvadia, 2025). While this applies across the public sector, the homelessness sector in particular should remain mindful of mistrust of institutions and people feeling disenfranchised, and consider this when rolling out and evaluating AI tools.[4]



What would count as evidence this tool is working?

Outcome measures for AI homelessness evaluations

Outcome measure selection should be guided by your Theory of Change. Many AI evaluations and case studies (e.g. Local Government Association, 2025a) focus on immediate measures like user satisfaction, or self-reported measures such as hours saved or user knowledge of housing rights. While user satisfaction is important to measure, perhaps to understand underlying mechanisms, it is usually not the main desired outcome; rather, we want to understand whether the service improves outcomes for users. Self-reported measures on time savings can be useful but these are subject to bias and imperfect recall.

For example, a chatbot like Housing Helper could successfully resolve many queries with accurate advice, but if these users are just getting faster advice without this improving their housing stability, then the tool hasn't delivered real value.

Local authorities should shift their focus away from operational metrics as evidence of impact and towards concrete outcome measures.

Measures

The outcomes we would first consider are: (1) homelessness duties owed (prevention, relief, and main duties, recorded in every local authority's statutory data), because they are the closest (although imperfect) administrative measure of whether homelessness occurred; (2) sustained tenancies, where your data allows you to observe them; and (3) contacts with the council's services, as a measure of engagement and early help-seeking - (see, however, the note on ambiguous direction of effect below). Satisfaction, usage, and self-reported confidence belong in your monitoring and process evaluation, rather than as your headline outcome.

Local authorities are required to report their homelessness activities by completing the quarterly Homelessness Case Level Information Collection (H-CLIC) statistical return. The H-CLIC database includes data on:

- The demographics of people making homelessness applications
- Applicants' support needs
- Circumstances and causes of homelessness
- The local authorities' decisions over their applications, including how many prevention/relief/main housing duties they owe
- All interventions undertaken by local authorities, including details on any temporary accommodation offered

H-CLIC data can be accessed in aggregate on the GOV.UK website, and approved researchers can access the data via the UK Data Service.

Local authorities also have their own homelessness datasets which you may be able to access. Some authorities subscribe to household-level data platforms built on benefits and arrears data - such as Policy In Practice's Low Income Family Tracker (LIFT) dashboard - which can be used both to identify an at-risk population and to observe changes in household circumstances. In the Housing Helper trial, Southwark Council's LIFT dashboard was how we identified the approximately 9,000 at-risk households in our sample.

The Centre for Homelessness Impact (CHI) has published a review of survey outcome measures in domains such as mental health, financial stability, quality of life, integration into the community, and attitudes towards job-seeking. This resource can be used to confidently select outcome measures.

Ambiguous direction of effect - Setting up measures

There may be some ambiguity as to the direction of the effect that an AI homelessness tool will have, depending on your outcome measure. Consider staff time spent on housing support as an example: providing more timely, relevant information could reduce the housing support a person needs from their local authority. Alternatively, more information may mean that a person is empowered to get in contact with their local authority, and therefore the tool will increase the amount of support the person receives from their local authority (see also discussion of agentic flooding in System effects, below). In setting up your outcome measures, it is important to anticipate these potentially multi-directional effects, so that if both positive and negative effects are happening at the same time, they don't cancel each other out; this can lead to the false impression that the intervention has no effect. See Box 3 for how we handled this when evaluating Housing Helper.

Case Study - Housing Helper's Outcome Measures

For the Housing Helper evaluation, we used our Theory of Change to select outcome measures.

Primary outcome

The number of relief and main housing duties owed to those who are encouraged to use Housing Helper, compared with participants in the control group. We will be collecting this data during the intervention period and for one year after, as we expect there to be a delay between someone being encouraged to use Housing Helper and being owed a homelessness duty.

This was chosen as the primary outcome because the impact that Housing Helper is aiming to have is a reduction in users becoming homeless.

Secondary outcome

The difference in the number of prevention duties owed to those who are encouraged to use Housing Helper, compared with participants in the control group, measured for the period of the intervention plus one year.

This was chosen as a secondary outcome because being owed a prevention duty is a very good indication that a person is in a housing crisis but is not yet at the point of becoming homeless. This helps us to understand whether Housing Helper helps to prevent homelessness, but does not prevent people from going deep into a housing crisis, or whether it prevents people from reaching the point of needing a prevention duty in the first place.

While creating the Theory of Change, we identified an ambiguity as to the direction of the effect that Housing Helper will have, with sound arguments for it either increasing or decreasing the number of homelessness duties owed by Southwark Council. Housing Helper may increase the total number of homelessness duties if using Housing Helper leads to a greater number of households understanding their housing rights and the Council's duty to help prevent them from experiencing homelessness. However, it could also result in a lower total number of homelessness duties, as Housing Helper may provide users with the knowledge and support they need to resolve any housing issues they may have before they reach the point where they would be owed a homelessness duty.

Relief and main housing duties are expected to decrease regardless of whether the total number of homelessness duties increases or decreases because we theorise that as a result of using Housing Helper, more people will get help with their housing earlier, and therefore not be in need of the very involved and serious intervention of a relief or main housing duty. However, prevention duties could move in either direction. Other research questions will be answered in the implementation and process evaluation.

Time lag between intervention and outcome

Consider when the best time to collect data will be. If you expect there will be a delay between the administration of the intervention and an outcome, consider delaying data collection or collecting data over an extended period of time after the intervention has taken place. This is why the Housing Helper trial collects outcome data for a full year after the intervention window closes, as a household helped (or not helped) by advice today may not appear in duty statistics for many months. For a commissioner, the practical implication is both budgetary and contractual - if you are to test homelessness outcomes, the evaluation must be funded and the data access must remain in place well beyond the point of the tool's launch.

Subgroup analysis

It is important to consider whether the intervention at hand might work differently for different groups of people, and ensure that you are including sufficient numbers of those demographics in your sample, that you are collecting appropriate demographic data, and that you are conducting subgroup analyses accordingly. It is likely that you will have considered which subgroups are the most important for your evaluation when creating the Theory of Change. Differential impacts are a particular concern for AI tools, which can perform differently across languages, ages, and levels of digital confidence.

Two of the decisions that determine whether you can answer this question are made at the early stages of the evaluation. The first is who you draw your sample from. The households who are most at risk – people without a settled address, without a smartphone or reliable data, with limited English or low literacy – are often the least well captured in the data local authorities use to identify an at-risk population. If a group is missing from that starting list, it will be missing from your findings.

The second is how many people you have. The Magenta Book AI Annex notes that with AI, variation in impact may be unusually pronounced and hard to anticipate, and advises deliberately boosting the numbers for groups that are poorly represented in the target population (Evaluation

Task Force & HM Treasury, 2026). This matters more in homelessness than in most fields, because statutory duties are relatively rare events, and dividing an already small number of them between groups can mean small samples with uncertain answers. It is important to be selective in choosing your subgroups: a small number of groups, chosen because your Theory of Change gives a reason to expect a difference, will be more valuable than a long list of characteristics not grounded in theory.[5]

Case Study - Running subgroup analyses for the Housing Helper evaluation

For the Housing Helper evaluation, we will be conducting separate analyses to see if the tool has differential effects on the following groups:

- **Age** - we expect that younger users may differ in their outcomes from older users on average, due to familiarity and comfort with AI tools and talking on the phone/in-person to someone who can give housing advice.
- **Rental/council tax arrears** - we theorise that those in payment arrears may benefit more from Housing Helper advice, compared with those with more complex challenges.
- **Immigration background/English as an additional language (EAL)** - we expect that those with an immigration background/ EAL status may particularly benefit from Housing Helper's ability to give housing information and advice in their preferred language.

How do I run a fair test of the AI tool?

Further considerations when designing a randomised trial of an AI tool in homelessness

Up to this point, this guide has set out principles that apply to anyone commissioning or evaluating an AI homelessness tool. This section turns to the specifics of designing a randomised trial, the most reliable way of identifying the impact of an intervention (Sanders et al., 2025).[6] These design choices – who gets encouraged to use the tool, what the control group receives, what counts as "use", when a trial should stop – determined the design of the Housing Helper evaluation, and they are likely to arise in trials of similar tools.

If you are commissioning rather than designing an evaluation, Box 5 summarises the key points. As the detail of these questions is mainly for evaluation teams, we cover them in **Appendix B**, using the Housing Helper trial as a worked example. Those who want the full technical detail – including sample size calculations, randomisation procedures, and the analytical strategy – should consult the **Housing Helper trial protocol**.

Trial design - key takeaways for commissioners

1. **You can randomise even when you can't mandate use** - trials of tools like chatbots typically randomise the invitation to use the tool (an "encouragement design"), which suits digital tools well. Note, however, that low take-up will reduce your study's statistical power, and you will need to adjust your required sample size accordingly (see Appendix B).

2. **The control group won't get "nothing",** as people are legally entitled to homelessness support, so the trial tests whether the tool is more effective than existing services.

3. **"Using the tool" needs a definition agreed up front.** Measuring it requires access to the tool's usage data.

4. **Trials can have agreed stopping points ("gateways").** Pre-set thresholds, such as minimum uptake, that stop the evaluation early and protect your budget if it looks like the tool isn't being used.

5. **Ask whether the tool works differently for different groups.** AI tools can perform unevenly across languages, ages and levels of digital confidence, and this needs demographic data collected from the start.

What if running a randomised controlled trial is not feasible or proportionate?

Running an RCT can be time- and resource-heavy, and not all AI tools can be, or should be, evaluated in this way. This may be because the nature of the tool is not conducive to recruiting an adequate sample size, or because the events that the AI tool is meant to prevent are exceedingly rare, leading to a sparse data problem. A trial may also be disproportionate, and a small pilot or modest spend on a low-risk intervention may not justify a full experimental evaluation (see Box 6). Furthermore, your authority may not have capacity or budget to support a trial. Although we strongly encourage conducting an RCT, if it is not feasible, or not proportionate, there are alternative methods that can be used which will be able to generate some evidence as to whether the tool is effective.

Proportionality and evaluating AI homelessness tools

The Magenta Book highlights that different interventions require different levels of evaluation, and that more substantial evaluations are appropriate where interventions are high cost, high risk, high profile, where there is high uncertainty about their effects, or where there is high learning potential (Evaluation Task Force & HM Treasury, 2026). It is worth applying these criteria to AI homelessness tools, because as the guidance sets out, they will generally meet several of them.

Applications of AI to homelessness are higher risk than in many other domains - the damage of a chatbot giving someone confident but incorrect advice about an eviction notice, or failing to make a safeguarding referral, could be severe.

There is also high uncertainty - while many promising approaches are being piloted, the evidence base on AI in the public sector is thin, and the evidence base on AI in homelessness is even thinner.

These tools are also high profile, with national policy strongly encouraging AI adoption, and local authorities can expect scrutiny of these deployments.

AI tools also tend to cost more than their initial price suggests - beyond licensing, local authorities face costs in staff time, integration, and ongoing maintenance, as well as risks of vendor lock-in.

This is also why the learning potential is high. Local authorities face the same statutory duties, are under similar pressures, and are considering similar tools, so one robust evaluation can help answer a question for the whole sector.

Proportionality remains an important consideration, and this guide sets out what to do when a trial is not the right approach. However, we anticipate that AI homelessness tools will usually warrant more evaluation rather than less, and that where an RCT is feasible, it will often be the appropriate choice.

Proportionality is usually a question of how much evaluation an intervention warrants. It is also worth asking what the decision you need to make requires, and when you need to make it.

- Deciding whether the tool is safe to deploy at all needs assurance from the developer and your own pre-launch testing, and comes before any resident is given access to an AI tool.
- Deciding whether the tool is safe to keep running needs ongoing accuracy checks and safety data, and you should be able to make this assessment within weeks or months of launch and continuously thereafter.
- Deciding whether it is reaching the people it was intended to help needs usage and demographic data, and may take a few months.
- Deciding whether the tool reduces homelessness duties, and whether to scale it, needs an impact evaluation, and in homelessness that will likely mean taking the intervention period plus at least a year, to allow enough time to detect an impact.

The first three of these (safe to deploy, safe to keep running, and whether it is reaching people) are the minimum every local authority should be concerned with. More details about these minimum requirements can be found below, in 'The minimum for every local authority'.

Thinking about decisions this way also helps clarify what it means to adjust quickly. You can change how a tool is implemented – the encouragement messaging, the referral routes, the prompts and sources behind the tool itself – relatively quickly, and your evaluation should be designed to allow this, provided changes are logged so that you can still interpret your results (see 'Intervention fidelity'). You can also agree in advance that the evaluation will stop if the tool is not being used, which protects your budget and your staff's time (see 'Gateways' in Appendix B). What you cannot speed up is establishing whether the tool works. Adjusting delivery during a trial is good practice, but treating early usage or satisfaction figures as evidence of effect, and scaling on that basis, should be avoided.

Quasi-experimental designs (QEDs)

QEDs can be used to estimate the causal effect of an intervention when the random assignment of participants to treatment and control groups is not possible. QEDs can be useful if the intervention that you are evaluating has already been implemented, or if you cannot assign participants to treatment or control groups for ethical or practical reasons. QEDs typically have poorer internal validity than RCTs, which means that you cannot be as certain of a trustworthy cause and effect relationship between the intervention and the outcome as you can be when you have conducted an RCT.

The Magenta Book AI Annex highlights that, as with experimental designs, the digital nature of AI tools means that there can be opportunities for QEDs (Evaluation Task Force & HM Treasury, 2026). For example, where a pre-determined cut-off on a continuous score determines who uses or is affected by the AI tool (an approach that may be applicable to the use of predictive analytics in homelessness), this opens up the opportunity for a regression discontinuity design. Alternatively, if an intervention is staggered, this may enable the use of difference-in-differences.

Experimental and quasi-experimental approaches are important for making credible causal claims about impact. A simpler design, such as comparing a group of people who received the intervention with a group who did not,creates selection effects, because the two groups are likely to differ from each other in meaningful ways, and so any difference in their outcomes cannot be confidently attributed to the tool.

For more on alternative approaches to evaluating AI interventions where a randomised trial is not feasible, including using theory-based methods, see the Magenta Book AI Annex. For more on choosing the right evaluation design, and on analytical methods generally, see the Magenta Book and the Magenta Book Analytical Methods Annex.

The minimum for every local authority

Whatever evaluation design you use – including if you are running a full RCT – there is a minimum that every local authority should have in place for any AI homelessness tool it deploys.

Two of these essentials come before launch, and have been covered earlier in this guide. The first is a Theory of Change. If you cannot articulate a credible pathway from the tool to the outcomes you want, you should not be deploying it (see 'What problem is the AI tool solving, and is AI the right answer?'). The second is assurance that the tool is safe to use and can perform the essential tasks it is designed for - the model evaluation discussed in 'What does this guide cover, and what does it not?'. Neither of these requirements needs specialist evaluation capacity, and they should both be considered before the deployment of any AI tool.

Once the tool is live, there is a minimum set of monitoring data that every local authority should collect. This need not be onerous. The Magenta Book AI Annex notes, the digital nature of AI interventions means much of this monitoring data can be collected automatically, providing early warning of any unintended consequences (Evaluation Task Force & HM Treasury, 2026). At a minimum, collect:

- **Usage** - who is using the tool, how often, and where in the journey people stop using it. Without this you cannot know whether the tool is reaching anyone at all, and low or falling usage is often an early sign that something isn't working.
- **Accuracy** - ongoing human-in-the-loop checks that the tool's outputs (whether advice, triage decisions, or drafted documents) are correct and current, alongside a vendor change log recording updates to the model, prompts, and sources (see 'Is the tool being implemented as intended?'). This matters because both the tool and the wider housing landscape of the local authority it operates in will change over time.
- **Safety** - safety incidents, escalations, and, where applicable, safeguarding referrals, including whether referrals were appropriate and actioned.
- **Reach and equity** - an extension of usage. It is important to know the key characteristics of who the tool isn't reaching. AI tools can perform unevenly across groups, and people at risk of homelessness are disproportionately digitally excluded, so a tool can look successful in aggregate while failing the people who may benefit most from it.

Collecting this data from the outset also preserves your options for a later impact evaluation. However, as set out above, this should usually be treated as the bare minimum, rather than sufficient. Monitoring data can tell us whether the tool is being used safely, but it cannot tell us whether the tool works.

Is the tool being implemented as intended?

Implementation and process evaluations: working out how the AI tool is being used in practice, and how it creates outcomes

An implementation and process evaluation (IPE) focuses on understanding how an intervention is put into practice, how it operates to achieve its intended outcomes, and the factors that influence these processes. IPEs complement RCTs, QEDs, or whatever other impact evaluation method you are using. The Centre for Homelessness Impact has published **IPE guidance** for homelessness evaluations, and TASO has a **range of useful resources** available on running an IPE.

There are specific things to consider when planning an IPE for an evaluation of AI tools in the homelessness sector:

Intervention fidelity

It is important to understand whether the intervention was delivered as it was intended. This means that you will know what is causing an effect if you find there is one. Both AI models themselves and the landscape they operate in are constantly changing. We want the AI tool to be giving the best possible service to users at all times, so it is likely for the intervention to need to change over the course of a trial, in response to the changing housing landscape. If this is the case, it is important that any updates or changes in the AI model, prompts, or retrieval augmented generation (RAG) sources are logged so that you can see if one of the changes made the tool produce different kinds, or different sizes, of outcome for users (see Box 7).

If applicable, the AI tool should be monitored to make sure it is giving accurate information to people and that it is not hallucinating. This is particularly important in homelessness services where serious harm could be caused by giving someone incorrect information. For Housing Helper, the risk lies in the advice itself - a question around an eviction notice is the kind where a confident but incorrect answer could cause serious harm. Before launching a trial, the AI tool should be tested for accuracy by people with expertise in housing or the field relevant to the tool, and where possible by people with lived experience of homelessness, to ensure the tool is accurate and useful. During the trial, tool outputs should be monitored by a human to ensure the tool stays accurate.

Case Study - Monitoring intervention fidelity for the Housing Helper evaluation

For the Housing Helper evaluation, Change Please kept a log of any changes they made to the model.

To capture passive changes, the research team continually tested Housing Helper and monitored chat logs for advice accuracy. We were aware of policy changes, such as the introduction of the Renters' Rights Act, which might lead to Housing Helper giving inaccurate advice, and we paid particularly close attention during this time.

Housing Helper's design also builds accuracy safeguards in, which are a useful benchmark for what to require of a vendor. The tool runs on a retrieval-augmented generation (RAG) system – it only generates answers that cite passages retrieved from approved sources (such as GOV.UK and Shelter). The model is set to be deterministic (a 'temperature' of zero), so it is more likely to report what the sources say rather than generating plausible-sounding variations. Where no approved evidence is retrieved, it declines to answer and signposts official links instead. Before launch, the tool was tested by Southwark housing officers, the evaluation team, and CHI's lived experience associates. A vendor should be able to describe these kinds of safeguards, and how they are maintained over the life of the tool.

Understanding engagement and uptake

You can utilise qualitative methods in the IPE to understand why people did or did not engage with the AI tool. Some possible reasons relevant to the homelessness sector might be language barriers, trust in AI, and prior interactions with the council. This is particularly important for an encouragement design, such as the Housing Helper trial. The IPE can help you understand whether the tool is failing to reach the people it was bought to help.

System effects

An AI tool is deployed into a service with finite capacity, rather than in isolation, and some of its most important effects may show up elsewhere in that service rather than in the tool's own metrics. For example, a tool like Housing Helper that empowers residents to seek help earlier could substantially increase requests for prevention duty assessments, leading to bottlenecks and growing waiting times for support. A tool can also change the composition of demand, as better-prepared residents arriving with drafted letters and documented cases may make casework faster, while also revealing complex cases that were not previously identified.

System effects from the use of more general-purpose AI tools are already being observed. Councils are reporting a steep increase in AI-assisted complaints and information requests, referred to as "agentic flooding" (The Economist, 2026). As well as an increase in volume, requests are reportedly also much lengthier, citing legislation and regulations, sometimes incorrectly, potentially leading to slower responses (Vallance & McMahon, 2026). Whether an increase in the volume and complexity of complaints means success or failure for a given AI tool may depend on whether the service receiving the requests can absorb them.

The IPE is where you find this out, which means looking beyond the tool's users to the whole service context, and speaking to the housing officers who receive referrals about referral volumes, referral quality, and workload. As the Ada Lovelace Institute has argued, evaluations of public sector AI often only assess the tool's own performance, missing the broader service and social context the tool is deployed into, and the need to keep evaluating iteratively as the tool and the service adapt to each other (Parker et al., 2025). In the Housing Helper evaluation, this is one reason the IPE includes interviews with Southwark housing officers, not only tool users.

Is the tool worth what it costs, and what costs and benefits could I be missing?

To determine if the homelessness AI tool is worth what it costs, you should run an economic evaluation. Much of this process will be similar to conducting an economic evaluation for other homelessness interventions. For more information on how to run an economic evaluation, please see the UK government's **Green Book** and the **Greater Manchester Cost Benefit Analysis (GMCBA) model**. CHI's **Economic Evaluation Guidance** and corresponding **Relevant Sources** outline recommended approaches to economic evaluation in homelessness. Costs and benefits should be counted from both a public sector and a societal perspective, and disaggregated by stakeholder - the council, the user, and wider public services.[7]

In addition to the approaches set out in these broad guidance documents, there are some considerations for the economic evaluation that are specific to AI tools, discussed below.

AI-specific costs

An economic evaluation of AI-enabled interventions should account for a range of costs that are often underestimated by local authorities. Beyond licensing fees, substantial staff time may be required to set up, train on, and maintain the tool, as well as to support its integration with existing systems and workflows. Ongoing costs can also arise from maintaining prompts and retrieval-augmented generation (RAG) components to ensure outputs stay accurate.

There are also possible costs associated with vendor lock-in, where dependence on a particular supplier limits future flexibility. There is also a Software-as-a-Service (SaaS) risk that subscription costs may escalate over time or that a supplier may cease trading or withdraw support after the tool has become embedded in routine practice, creating disruption and replacement costs. A recent House of Commons committee report highlighted that the public sector as a whole is "dependent on a small number of large digital and technology providers… and the government appears to be worryingly comfortable with this" (Science, Innovation and Technology Committee, 2026).

It is also important to acknowledge negative externalities associated with AI-enabled interventions, such as environmental costs related to computing infrastructure and concerns about poor labour conditions within the AI supply chain. For more information on the impact of AI on the environment, please see this blog post by the UN Environment Programme. For more information on negative externalities associated with labour markets, collection of personal data, and democracy, please see Acemoglu (2021). The International AI Safety Report provides a wider evidence base across these areas.

These costs can be difficult to bring into an economic evaluation, as they are incurred in the supply chain, before the point of purchase, and are not reflected in what a council pays for a licence. However, that does not place them outside the scope altogether. Local authorities are already required, under the Public Services (Social Value) Act 2012, to have regard to economic, social and environmental wellbeing when procuring services (see the Ada Lovelace Institute's Buying AI, and Who pays for AI risks? on how the costs of AI harms are currently distributed). We would not suggest that councils attempt to quantify all of these effects, but we would suggest that they are named, and asked about, at commissioning stage.

Ambiguous direction of effect - Economic evaluation

As discussed above in Box 3, homelessness prevention interventions have an ambiguous direction of effect on the number of homelessness duties owed.

When you have analysed the data you have collected and know the size and direction of the intervention's effect on homelessness duties, then you can work out the cost to councils. This matters for the economic case as, if the tool increases prevention duties in the short term, the council's costs may rise even where users benefit, so the perspective (council vs. societal) and time horizon of the economic evaluation need to be set with this in mind.

Gateways for economic evaluation

You can also establish a gateway for conducting an economic evaluation, where you do not conduct an economic evaluation if you do not find a statistically significant effect. In the Housing Helper trial, we pre-specified that the economic evaluation would only go ahead if the impact evaluation found evidence of an effect on homelessness duties. Without clear evidence of benefits, there is not much an economic evaluation can add. If there are only costs, and no evidence of benefit to users, the recommendation is clear: do not implement the intervention.

What could go wrong, and what do I need in place before I commission?

Data, ethics, and governance: the AI-specific bits

There are specific measures that should be in place before running an evaluation of a new intervention. While we set out some points relevant to homelessness here, you should draw on further resources to ensure your approach is ethical and legal.

Please see the **Data and AI Ethics Framework** for more information on responsible development, procurement and use of data and AI in the public sector. For data protection specifically, the **ICO's guidance on AI and data protection** will likely be a key source, and you should expect to complete a Data Protection Impact Assessment (DPIA) before deployment. You may also wish to consider publishing a record for the tool under the **Algorithmic Transparency Recording Standard** (ATRS), which is mandatory for central government and strongly encouraged for local government.

Treat chatbot logs as sensitive data

Chatbot logs for homelessness tools should be treated as sensitive data, because it is likely that the logs will contain safeguarding disclosures, information about someone's immigration status, and health data. This has implications for data minimisation, retention, and access.

- **Data minimisation** - for evaluation purposes, only a small sample of chat logs needs to be seen by a researcher, in order to understand how people usually use the chatbot and to understand what types of conversations trigger the chatbot to refer an individual to other services.
- **Data retention** - sensitive data should not be kept for longer than you need to use it for. All copies of chat logs held for evaluation should be deleted after analysis by a researcher.
- **Data access** - only authorised researchers who are actively analysing the chat logs should have access to them.

You cannot evaluate an AI tool effectively if you do not collect the right data, but you should not collect more data than you need. Commissioners, evaluators and vendors should discuss exactly what the tool will log, for what purpose, and for how long.

Safeguarding and referrals

If applicable to the AI tool you are evaluating, systems for safeguarding and referrals should be devised and thoroughly tested before launch. All parties should understand what will trigger a referral, and what will happen to the individual when the referral is made. The appropriate checks should be made to ensure that the agreed and tested processes are happening. For a tool like Housing Helper, this means making sure a safeguarding officer can review chat logs in order to establish that the AI tool is triaging users properly .

Case Study - Safeguarding and referrals for Housing Helper

Housing Helper triages users into different risk categories based on their answers to questions the chatbot poses. The response is proportionate to the level of risk.

For the Housing Helper evaluation, all parties had to agree on the best way to keep users safe, and balance our duty to refer with users' data protection rights. This was a difficult balance to get right, and involved many conversations between Southwark Council, Change Please, and the Centre for Homelessness Impact's Data Protection Officers, lawyers, and Safeguarding Leads.

The process that was decided on is shown in Appendix C.

How do I share what I've learned, and learn from what others are doing?

Transparency, learning and collaboration

At the time of writing, there is limited evidence on the efficacy of AI tools in homelessness, as well as the public sector more widely. When your evaluation is complete, sharing the results widely so others can learn from your work will help other local authorities to make better decisions with their funding, and to tackle homelessness more efficiently. We would encourage all local authorities who have completed an evaluation to add their evaluation to the Local Government Association's **Artificial intelligence case study bank**. It is important to share the results of your evaluation regardless of what you find. Negative or null results are just as useful as positive results, because they allow us to learn what doesn't work and in doing so prevents us from wasting valuable resources.

Box 9: Case Study - Sharing findings

Findings from the Housing Helper trial will be published regardless of whether results are positive or negative. You can find the **project page** on CHI's website, as well as the **trial protocol**, with full details about the trial. For more detail on CHI's approach to transparency and learning in evaluation, see CHI's **Evaluation Pillars**.

Collaboration can also start earlier than publication. If your authority is considering a tool that another council is already testing, or if you face the same delivery challenge, talk to other councils before you buy. Councils using similar tools can share due diligence, align on outcome measures, and in some cases run joint evaluations, pooling participants across local authorities.

Box 10: Considering the case for a What Works Centre for AI in public services

In their policy briefing **Measuring Up** and other **related pieces**, the Ada Lovelace Institute outlines a case for establishing a dedicated What Works Centre for AI in Public Services (Deshmukh & Davies, 2026; Parker et al., 2025; Parker, 2026). Their argument is in line with what we have set out in this guidance – that there is strong enthusiasm in the public sector for AI, but that much of this is based on short-term metrics, while neglecting long-term, concrete outcomes, trade-offs, and the perspectives of frontline workers. A centralised, independent What Works Centre could help build the evidence base to address this. Such a centre could commission and synthesise independent research, and host a shared repository of public sector AI use cases to reduce duplication.

The rationale for a standalone centre, rather than relying on existing, sector-specific What Works Centres, is the expertise required for evaluating AI interventions. What Works Centres may not always have the necessary specialised technical and data-science capabilities in-house that would ideally be used in evaluating AI interventions. An AI What Works Centre would enable local authorities to confidently weigh the opportunity costs of AI against alternative interventions, ensuring public funds deliver value.

Closing recommendations

Local authorities are adopting AI tools at pace, and adoption is outpacing evidence. In homelessness, the harm from deploying AI badly can be severe, and this is one reason why it is so important to test before we scale. Decisions to procure, scale up, or discontinue these tools should be informed by evidence of effectiveness, safety and value for money, rather than by technological novelty or anticipated efficiencies alone.

Before you buy, ask whether AI is the right answer. Carefully consider the value an AI tool adds on top of the housing and homelessness services you already have available. If you cannot articulate a credible Theory of Change for the tool, any attempt at deploying or evaluating it should stop there.

Decide what would count as evidence that the tool is working, and measure outcomes that matter – homelessness duties, sustained tenancies, contacts with the council – not just user satisfaction and hours saved. Consider also whether the tool has potential for ambiguous direction of effects for your chosen outcome measures, and determine whether multiple outcome measures could be necessary for disentangling the tool's impact. Articulating your Theory of Change will help with this.

Then aim for the strongest evaluation design you can support. The digital nature of these tools makes randomisation easier than you might expect, and an implementation and process evaluation alongside it will tell you how the tool is being delivered and why it succeeds or fails. Importantly, do not rely on self-selection to create your treatment and control groups – those who elect to use AI will almost certainly be meaningfully different from those who do not. Where a trial is not feasible or proportionate, use an alternative design such as a quasi-experimental design, and be transparent about its limitations. At a minimum, put in place what every local authority needs: a Theory of Change, assurance the tool is safe, and the minimum monitoring data on usage, accuracy, safety and reach.

While the tool is running, expect it to change mid-evaluation. Evaluation should be an ongoing process, not a single assessment conducted before the tool is rolled out -keep a meticulous change log and ongoing accuracy monitoring, particularly given that statutory obligations and best

practice change over time, and make sure that safeguarding procedures remain effective as the tool changes. Treat chat logs as sensitive data, and require human-in-the-loop escalation for high-risk cases. Expect a lag, too, as the outcomes that matter can take a year or more to appear in the data, so plan and budget for follow-up.

When the results come in, count the full costs and benefits from both a public sector and a societal perspective, including staff time and integration, what a prevented homelessness duty saves your authority in temporary accommodation spending, and the value of improved stability and wellbeing to users.

And share what you find. Negative and null results are just as useful as positive ones, because they allow us all to learn what doesn't work. Local authorities face the same statutory duties, are under similar pressures, and are considering similar tools –a robust evaluation can help answer a question for the whole sector. Greater transparency strengthens the collective evidence base and helps ensure that future investment in AI-enabled public services is guided by robust evidence of what works.

References

Acemoglu, D. (2021). Harms of AI. MIT Economics. URL: https://economics.mit.edu/sites/default/files/publications/Harms%20of%20AI.pdf (Accessed 21/07/2026).

Department for Science, Innovation and Technology. (2025). AI Opportunities Action Plan. URL: https://www.gov.uk/government/publications/ai-opportunities-action-plan/ai-opportunities-action-plan (Accessed 21/07/2026).

Deshmukh, S. & Davies, M. (2026). Measuring Up. Ada Lovelace Institute. URL: https://www.adalovelaceinstitute.org/policy-briefing/measuring-up/ (Accessed 21/07/2026).

Evaluation Task Force & HM Treasury. (2026). The Magenta Book. HM Treasury. URL: https://www.gov.uk/government/publications/the-magenta-book (Accessed 21/07/2026).

Hauser, O. P., Light, M., Shelmerdine, L., & Blumenau, J. (2025). Why evaluating the impact of AI needs to start now. Nature, 643(8073), 910-912. https://doi.org/10.1038/d41586-025-02266-7 (Accessed 3/09/2026).

Homeless Link. (2026). Artificial Intelligence and Homelessness. URL: https://homeless.org.uk/areas-of-expertise/improving-homelessness-services/supporting-innovation-in-homelessness-services/artificial-intelligence-and-homelessness/ (Accessed 21/07/2026).

Local Government Association. (2025a). Kingston Council: Using AI in Adult Social Care administration. URL: https://www.local.gov.uk/case-studies/kingston-council-using-ai-adult-social-care-administration (Accessed 21/07/2026).

Local Government Association. (2025b). State of the Sector: Artificial Intelligence - 2025 update. URL: https://www.local.gov.uk/publications/state-sector-artificial-intelligence (Accessed 21/07/2026).

Ofcom. (2025). Exploring Digital Disadvantage. URL: https://www.ofcom.org.uk/siteassets/resources/documents/internet-based-services/technology/research-digital-disadvantage/exploring-digital-disadvantage-research-report.pdf?v=393692 (Accessed 21/07/2026).

Parker, I. (2026). Working out a What Works Centre for Public Sector AI. Global Government Forum. URL: https://www.globalgovernmentforum.com/working-out-a-what-works-centre-for-public-sector-ai/ (Accessed 21/07/2026).

Parker, I., et al. (2025). Learn Fast and Build Things. Ada Lovelace Institute. URL: https://www.adalovelaceinstitute.org/policy-briefing/public-sector-ai/ (Accessed 21/07/2026).

Polo, N. & Modhvadia, R. (2025). Great (Public) Expectations. Ada Lovelace Institute. URL: https://www.adalovelaceinstitute.org/policy-briefing/great-expectations/ (Accessed 21/07/2026).

Sanders, M., et al. (2025). Designing and Delivering Randomised Trials for Social Policy. King's College London. URL: https://www.kcl.ac.uk/research/designing-and-delivering-randomised-trials-for-social-policy-1 (Accessed 02/09/2026)

Science, Innovation and Technology Committee. (2026). Rewiring the state: Delivering digital government. House of Commons. URL: https://publications.parliament.uk/pa/cm5902/cmselect/cmsctech/61/report.html (Accessed 21/07/2026).

Teixeira, L. (2022). Editorial: Introduction to the Special Issue - What Works: Research Impact and Homelessness. European Journal of Homelessness, Winter 2022. Centre for Homelessness Impact. URL: https://www.homelessnessimpact.org/publication/european-journal-of-homelessness-introduction-to-the-special-issue (Accessed 21/07/2026).

The Economist. (2026). How AI is breaking the British state. URL: https://www.economist.com/leaders/2026/08/06/how-ai-is-breaking-the-british-state (Accessed 26/08/2026).

Vallance, C. & McMahon, L. (2026). Surge in AI-assisted complaints putting extra strain on councils and schools. BBC News. URL: https://www.bbc.co.uk/news/articles/c2dk6wzjw23o (Accessed 26/08/2026).

Endnotes

1 "Generative AI is a subset of AI capable of generating text, images, video or other forms of output by using probabilistic models trained across one or more domains. These models learn from large amounts of specially-curated training data to discern and replicate complex patterns and structures. The output generated by these models mimics the characteristics learned from that training data, enabling many novel applications – ranging from personalised content generation, to advanced analysis and evaluation, and aiding creative processes." https://www.gov.uk/government/publications/ai-insights/ai-insights-generative-ai-html

2 For more on predictive risk modelling and homelessness, see CHI's Using Data to Prevent Homelessness trial.

3 Here we use Housing Helper as an example, considering whether users trust the AI information they're given, but this also applies to other AI interventions. If a machine learning algorithm identifies someone as at risk of homelessness, but the housing officer delivering support does not trust the algorithm, this will limit the impact of the tool. This is discussed in Ada Lovelace Institute's Critical Analytics report, and will be explored in CHI's forthcoming Using Data to Prevent Homelessness trial.

4 For more on trust, public institutions and the homelessness sector, see recommendations from MHCLG's lived experience forum, and this report on experiences of stigma and discrimination from KCL.

5 From a more technical perspective, subgroups should be pre-specified in the protocol, and reported as secondary or exploratory rather than headline findings, since trials are rarely powered to detect differential effects and each additional comparison raises the risk of a false positive. Where a particular group is a priority, it can be over-sampled or used as a stratification variable at randomisation.

6 This guide assumes some familiarity with randomised trials. Sanders et al. (2025) offers both an introduction for readers who want one and more detail on design choices discussed in this section and in Appendix B.

7 CHI's Economic Evaluation guidance recommends both perspectives, with benefits and costs disaggregated by stakeholder. The supplementary guidance on data sources sets out the human costs of homelessness alongside the fiscal ones, and gives valuations that can be applied to benefits accruing to users.

8 In many encouragement designs, you will also get participants in the control group who choose to take part in the treatment despite not being encouraged to do so. In these cases, you can still estimate the effect for a select group of participants, called "compliers." This analytical approach is often called a CACE analysis, or Complier Average Causal Effect. In the case of Housing Helper, it is a new tool that has not been promoted publicly yet, so we do not anticipate that we will have many non-compliers in the control group, i.e. people in the control who use Housing Helper.

Appendix A:

Theory of Change

Situation: Local authorities have a duty to offer advice related to housing. The number of people approaching local authorities for housing-related advice and assistance is increasing and putting pressure on LAs.
This is particularly true when people approach the local authority when they are in crisis, at which point interventions are less effective and more costly. There may be opportunities to intervene early to prevent people from reaching a crisis point.

Aims: For low-risk individuals, offer targeted advice/assistance earlier than they would typically receive it to divert them from the Local Authority. Advice will be given on housing rights and the support people are entitled to, while assistance will be given requesting rent reductions and payment flexibility, requesting payment plans for debts, and identifying incorrect Section 8/Section 21 notices

- **For high-risk individuals, help them to receive support from the Local Authority more quickly by effectively triaging them and referring them to the LA**
- **Reduce the number of prevention and relief duties across all user types**
- **Increase the number of people who challenge their eviction notices across all user types**
- **Increase the number of people who take up support – from the council or other organisations**
- **Ensure that the advice is accessible to as many people as possible, particularly to people in arrears, survivors of domestic abuse, people with English as an additional language, care leavers**
- **Reduce time spent and costs for Local Authorities**

Process			Impact	
Inputs	**Activities**	**Outputs**	**Outcomes**	**Impact**
• LA knowledge of potential users • Retrieval Augmented Generation (RAG) system • AI training materials (websites including Citizens Advice, GOV.UK, Shelter, Southwark Council; Section 21/Section 8 notices) • AI persona development • Framework for continuously monitoring CPAI to ensure efficacy and safety • Risk categorisation framework for triaging users • Language translations • Referral pathway mechanism with LA	• Building CPAI • Preparing LAs to respond to high-risk users • LA sends an email and text messages to users to get them signed up • Change Please sends emails and text messages to users to encourage them to use the tool or return to the tool • Encouraging people to use CPAI (A/B testing messages) • Building dashboard for council use • Continuous monitoring of CPAI, comprising bias detection, stakeholder feedback, performance monitoring, and quarterly review of ethical compliance	**All users:** • Users respond to engagement messages and reminders, clicking the link to access the tool • Users create an account • Users are risk-triaged and high-risk users are referred to the LA to receive appropriate support • Users engage with CPAI meaningfully, on multiple occasions to revisit resources **Low to medium-risk users:** • Users are accessing customised, accessible, understandable, and useful housing advice • Section 21/Section 8 notices are uploaded to CPAI, custom advice and populated response letters are produced • Users are sending letters to landlords asking for rent reductions/payment plans • Users with English as a second language can interact with the tool in other languages, which the tool can use accurately and naturally • CPAI connects users to local organisations • Users use other features offered by CPAI (e.g. benefit calculator tool) **High risk users:** • Users are flagged to a trained housing worker who is able to respond quickly to crisis	**All users:** • Users feel they are receiving high-quality and useful advice • Users access specialist services (e.g. domestic abuse/mental health services) • Users with English as an additional language confidently send and receive letters to landlords in English, with translations in their primary language • Users access advice and/or assistance earlier than they would have without using CPAI • Users feel comfortable engaging with CPAI • There is a differential impact across subgroups, particularly for age, literacy level, language, tenancy type **Low/medium risk users:** • Users are granted rent reductions • Users are granted payment plans for rent arrears, council tax arrears, other debts • Users approach the council/other organisations for assistance • Users challenge Section 21/Section 8 notices **High risk users:** • High-risk users receive timely support they need from the LA **LAs:** • LA time shifts from reacting to emergency cases, to time working with users before the crisis point **Possible unintended consequences:** • Users more likely to approach the council for assistance places strain on services • Users become over-reliant on CPAI • Users access non-council services over council services	• Fewer people become homeless • Fewer people are owed a relief duty or a main housing duty, saving the LA money • For lower-risk users, fewer people are owed a prevention duty, which saves money for the LA • For high-risk users, more people are owed a prevention duty because they accessed help earlier, saving the LA money in the long-term via fewer relief/main housing duties • LAs divert spending from crisis response to prevention

Rationale & assumptions

- The training materials used to develop CPAI are accurate
- CPAI will be able to give trustworthy advice and will not hallucinate
- The triage system effectively categorises users
- Encouragement messages will reach users

- Users remember CPAI when they need it
- Users have access to a phone/computer when they need to access CPAI
- Users will have the technical skills to effectively use the tool, including writing effective prompts and downloading/editing/saving documents

- People answer the questions asked by CPAI accurately
- People trust CPAI enough to use it and give it sensitive data
- Users spend enough time with the chatbot for tailored and helpful advice to be delivered

- People trust the information given by CPAI to action it
- Asking for a rent reduction/payment plan for debts results in users receiving the rent reduction/payment plan
- Rent reduction/flexibility, payment plans for debts, connection to services, and/pr additional support is sufficient to prevent homelessness

Appendix B: Trial design

This appendix is written mainly for evaluation teams, and for local authority staff who want to understand an evaluation proposal in more depth. It covers four design considerations that arise when running a randomised trial of a tool like an AI housing advice chatbot, using the Housing Helper trial as a worked example throughout. It is not a complete guide to trial design - for the general methods, see CHI's statistical guidance and KCL's Designing and Delivering Randomised Trials for Social Policy. For the full detail of the Housing Helper trial, including sample size calculations, randomisation procedures, and the analytical strategy, see the trial protocol. What we set out here are the design choices where evaluating an AI tool in homelessness differs from evaluating other interventions, and how we resolved them.

Encouragement designs

Instead of randomising who receives treatment, we randomise who will be encouraged to take part in treatment, effectively giving a boost to that group.[8] If the encouragement works, and more people in the encouragement group use the tool than in the control group, you can estimate the combined effect of the invitation to use and using the AI tool by comparing the average outcomes of the randomised groups. This helps to overcome selection bias when participants' choices must be accommodated. For more information on encouragement designs, please see this report from J-PAL on challenges to randomisation and their solutions. Encouragement designs are particularly helpful for evaluating AI tools in the homelessness sector, for a number of reasons: (1) it is impossible to mandate that people assigned to the treatment group use the AI tool, and it is highly likely that the majority of people in the treatment group will not use the tool; (2) there is likely to be substantial selection bias between people who would and wouldn't use an AI tool, particularly amongst people experiencing homelessness who already experience high levels of digital exclusion (Ofcom, 2025); (3) the digital nature of an AI tool means that you can very cleanly randomise people into the treatment and control groups.

Encouragement designs require some additional considerations, however. By their nature, encouragement designs need a larger sample in order to achieve similar levels of statistical power, compared with more conventional treatment allocation designs. This is because not everyone allocated to receive the encouragement will take up the treatment. In the case of AI tools, where the encouragement might take the form of an email or text message to sign up for a tool, you may find that only a small minority (5-10%) will in fact comply and receive the treatment. Therefore, you need to adjust your sample size accordingly, based on how successful you think your encouragement design will be in getting users. We recommend coming up with a few different scenarios for this, based on two different things: (1) what you know about your population of interest, e.g. how easy are they to contact?; (2) engagement rates on similar campaigns or industry averages, e.g. email open rates, click-through rates. For example, if we expect a 20% click-through rate, 40% account creation rate, and 75% meeting the use threshold for compliance, that's 0.2*0.4*0.75 = 0.06, or a 6% compliance rate. Use this value to inflate your required sample size. See Chapter 6. Sample Size and Power in Sanders et al, (2025) for guidance on how to do power calculations for different trial designs.

Active controls

Employing an active control means that people assigned to the control group do not get nothing, but get business-as-usual services rather than the intervention being tested. This means that instead of isolating the baseline effect of the intervention, we are assessing its efficacy relative to usual practice. When evaluating a tool for use in the homelessness sector, using an active control is important because people are legally entitled to help to prevent and relieve homelessness, so we want to test whether a novel intervention is more useful than business-as-usual practice, rather than testing whether the intervention is better than the "receive nothing" scenario, which (at least in theory) does not exist. This links back to our earlier question of "is AI the right answer?" – the non-AI alternative, such as a webpage, leaflet or phone line, is what the control group should receive.

Case Study - Housing Helper's Encouragement Design and Active Control

The Housing Helper trial has an encouragement design, where individual households were randomised to receive encouragement to use Housing Helper, and the remainder received an "active control" message, signposting them to existing resources. An example of an encouragement message and the message for the active control group are below.

Figure 2. An encouragement message sent by Southwark Council to the treatment group

HOW AI CAN FIX YOUR HOUSING ISSUES

We know that dealing with housing issues can be stressful, and waiting for an appointment isn't always an option when you need answers right away.

USE THIS FREE AI TOOL TO:

- Draft letters for you
- Check your rights
- Assess your risk of homelessness
- Get confidential, verified advice, 24/7

ACCESS THE AI ASSISTANT NOW

Figure 3. The email sent by Southwark Council to the control group

Hello,

Southwark Council offers housing support services to residents who may need them. You can find more information and advice on our website at https://www.southwark.gov.uk/housing.

Please get in touch if you need any further support.

Defining compliance

It is important to decide in advance what will count as a participant complying with an intervention. Not pre-defining compliance risks counting people as treated when they have not engaged meaningfully with the intervention, which will dilute any effect of the intervention you may find.

Case Study - Defining Compliance for the Housing Helper evaluation

As described in our Theory of Change, Housing Helper has many different uses, and many different pathways a user could be sent on, depending on their profile and needs. Therefore, we had to agree a definition for whether a user had been treated or not, so we could decide which participants' data should be included in the CACE analysis. We decided on the following criteria for compliance:

Clicking the link provided in an encouragement message

Completing the user sign up flow and creating a Housing Helper account

Using Housing Helper for a minimum of four minutes

Four minutes was suggested by Change Please as the minimum time required to be risk triaged by Housing Helper and for the user to derive value of some kind. This is based on their experience using Housing Helper in the field (outside the context of this trial and Southwark borough). While we are using this as the cut-off time for determining compliance, we will also explore higher cut-off times that likely correspond to deeper usage as sensitivity checks in the analysis.

Gateways

Incorporating gateways into your RCT design means that if a pre-set target is not reached, you will stop the evaluation. These are very useful when running an RCT in homelessness research when you are unsure of your ability to recruit participants or whether you will have enough participants complying with the treatment. You can learn more about gateways and the 'What Works Faster' approach here.

Case Study - Housing Helper's Gateways

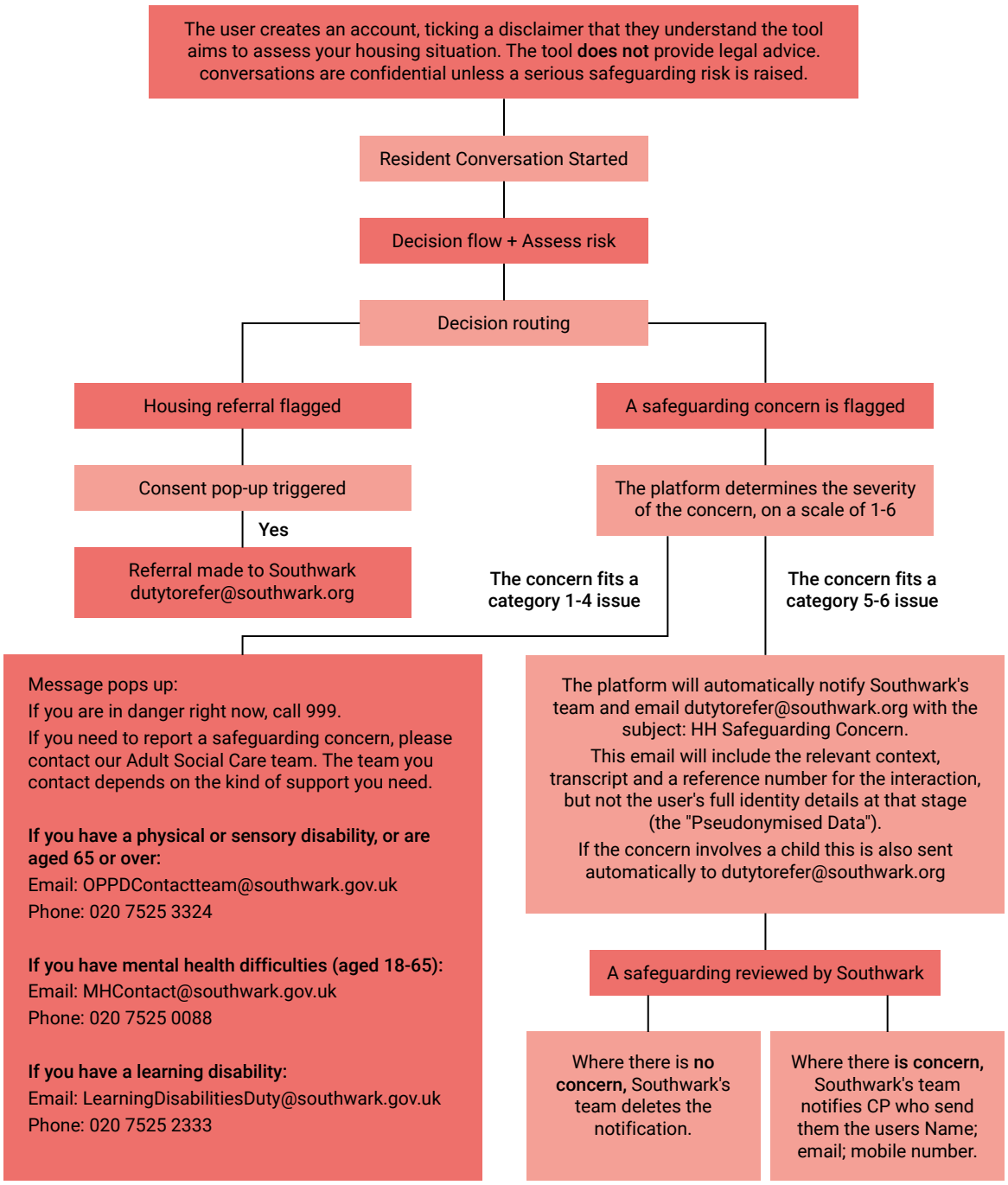
The compliance gateway

We needed to hit 7% compliance among the households randomised to receive encouragement, and would have ended the evaluation if we had not. The compliance gateway was set for two main reasons: (1) the tool is only valuable if we can persuade people to use it; if the public say (with their actions) that Housing Helper is not helpful, there is little point in testing its efficacy; and (2) we need a sufficiently powered study, and as discussed above under 'Encouragement designs', we had some uncertainty about expected rates of compliance in our sample. Setting a clear, agreed-upon threshold helped to ensure that if the encouragement campaign was less successful than hoped, we could save valuable trial resources and Southwark Council's time and efforts, and end the trial early.

The economic evaluation gateway

If Housing Helper has no statistically significant effect on homelessness duties, we will not run an economic evaluation. This is because without a statistically significant effect, there is not much one can say about the benefits of an intervention. If there are no benefits and only costs, an economic evaluation is not needed to recommend not implementing the intervention.

Appendix C: Safeguarding flow



Centre for Homelessness Impact

Centre for Homelessness Impact
www.homelessnessimpact.org

© 2026 | Centre for Homelessness Impact
ISBN 978-1-914132-62-9
CHI | Registered Charity Number: E&W1183026; SC049501.
Company Number: 11732500