

Society Under the Lens

Civil Society, Global
Institutions & Change.

A MORE
INCLUSIVE
SUSTAINABLE
JUST WORLD

PEOPLE
PEACE
PLANET
PROGRESS

PEOPLE

IDEAS
DIALOGUE

ISSUE 02 • SEPTEMBER 2026

BEFORE THE LIE TAKES HOLD

Prebunking Migration Misinformation and the Conditions of Social Cohesion

SINAN MERİÇ | AUTHOR & FOUNDER EDITOR

A research informed analysis, comparative evidence audit and behavioural pilot design

SOCIETY UNDER THE LENS

PEOPLE • POLICIES • PERSPECTIVES

OPEN ISSUE >

Abstract

Institutions usually enter the contest over migration after a claim has already travelled: a crime has been stripped of context, a benefits figure has been converted into a story of preferential treatment, or a local service failure has been attributed to an undifferentiated migrant population. This paper asks whether that sequence can be altered. It combines a research-informed analysis of misinformation and intergroup threat, a comparative audit of official evidence, and a field-ready design for testing fact-checking, prebunking and trusted-community-messenger interventions. The analysis rejects two convenient assumptions. The first is that every adverse claim about migration is necessarily false; the second, that a correct statistic is sufficient to change a social attitude. Official records from Sweden and the United Kingdom show why the first assumption is unsafe, while experimental work on corrections and inoculation demonstrates the weakness of the second. The resulting programme therefore separates factual accuracy from perceived threat, institutional trust, willingness to share and social distance. Its proposed *modus operandi* is diagnostic before it is persuasive: establish what the evidence permits, classify the claim, select an intervention appropriate to the audience and narrative, and test whether any immediate gain survives four weeks and three months. Prebunking warrants serious use, but not as an *ad hoc* substitute for credible policy, transparent data or the correction of genuine material pressures.

Executive findings

Question	Finding	Policy consequence
What need does the paper address?	Migration communication often begins only after a misleading narrative has acquired familiarity, emotional force and partisan ownership.	Build preventive capability while retaining rapid correction for claims already in circulation.
What does official evidence add?	It prevents advocacy from outrunning the record. The same dataset may disclose group differences, compositional effects and declining within-group rates at once.	Use an evidence verdict rather than a binary myth-versus-fact label.
What can prebunking plausibly do?	It can improve recognition of manipulation techniques and, in some settings, reduce willingness to share. Evidence for durable attitude change is thinner.	Measure epistemic and social outcomes separately and include delayed follow-up.
Why test trusted messengers?	A correct message from a distrusted institution may be discounted, while familiarity alone does not guarantee accuracy or fairness.	Treat messenger identity as an experimental variable, not a communications afterthought.
What is the paper’s original output?	A claim-audit protocol, an intervention decision matrix and a five-arm pilot design linking narrative, audience, messenger and outcome.	Give public institutions a reproducible basis for choosing and evaluating interventions.

Introduction

Migration misinformation seldom arrives as a self-contained false proposition. More often, it is assembled from a genuine pressure, an exceptional incident, a poorly defined administrative category and an insinuation of collective intent. A hospital waiting list may indeed have lengthened; the misleading step lies in treating migration as the sole cause without examining staffing, ageing, funding or regional capacity. A foreign-born suspect may appear in the police record; the distortion begins when suspicion becomes conviction, an individual becomes a population, and a change in absolute numbers is reported without reference to the size and composition of that population. The fact of the matter is that misinformation thrives less through complete fabrication than through the selective arrangement of what is partly true.

This circumstance places public authorities in an awkward position. Silence allows the first account to harden. A hurried rebuttal may repeat the claim, appear defensive, or ask citizens to trust the same institution whose legitimacy is already contested. A polished campaign may also evade an inconvenient reality: sometimes a community is experiencing a real distributional pressure, and no quantity of message testing can provide a school place or affordable dwelling. Communication policy consequently requires a prior act of classification. Before deciding how to respond, institutions must know whether a claim is unsupported, misleading, partly supported, context-dependent, supported or presently inconclusive.

The International Organization for Migration has placed misinformation and disinformation within its treatment of migration and human security, observing that distorted narratives can affect migrants as well as receiving communities¹. The European Parliament has likewise documented the capacity of migration-related disinformation to intensify polarisation and to instrumentalise existing anxieties². These diagnoses establish the seriousness of the problem, but they do not by themselves tell a municipality, ministry or civil-society organisation what to do on Monday morning. The present study is written for that decision.

Scope and method

The inquiry proceeds in three parts. Part I examines why migration claims are resistant to correction and what fact-checking, prebunking and trusted-messenger approaches can reasonably be expected to accomplish. Part II subjects prominent narratives to an evidence audit. It uses official records as primary evidence, then places those records beside peer-reviewed or methodologically explicit analysis. Part III translates the findings into a comparative pilot. The paper does not purport to estimate a universal causal effect of migration on crime, wages or public services. Such an undertaking would require harmonised microdata and research designs that exceed the purpose of a communication-policy study. Rather, it demonstrates the evidential discipline that should precede an intervention and proposes a way to test the intervention itself.

1. International Organization for Migration, World Migration Report 2024, chap. 5, ‘Migration and Human Security: Unpacking Myths and Examining New Realities and Responses’ (Geneva: IOM, 2024).
2. European Parliament, The Impact of Disinformation Campaigns about Migrants and Minority Groups in the EU, EXPO/IDA/2021/653641 (Brussels: European Parliament, 2021).

The source hierarchy is deliberate. Official statements are relevant because they reveal what governments claim and what categories they administer; they are not treated as dispositive merely by reason of their provenance. Administrative figures may depend upon reporting practice, enforcement priorities, legal definitions and denominator choice. They are therefore read with their methodological notes and, where possible, alongside survey evidence and adjusted research. This is particularly important in crime analysis, where recorded offences, suspected persons, prosecutions, convictions and victimisation are different objects. To move between them without warning is to manufacture certainty.

Part I Behavioural foundations and institutional limits

Migration misinformation as an explanation of insecurity

The persistence of a migration claim cannot be explained by ignorance alone. Hainmueller and Hopkins’ review of attitudes towards immigration found that citizens often respond to perceived national or sociotropic consequences rather than merely to direct personal competition.³ Intergroup-threat research adds that symbolic concerns about identity and realistic concerns about resources may work together, even when the immediate factual claim concerns only employment or welfare.⁴ A statement such as ‘migrants receive everything’ thus functions as more than a proposition about budgets. It offers a moral account of who belongs, who receives, and who has been neglected.

This is why accurate information can improve factual estimates without moving policy preferences by an equivalent degree. Hopkins, Sides and Citrin found that supplying correct information about the size and characteristics of immigrant populations reduced misperceptions but had little effect on immigration attitudes.⁵ More recent experimental work reaches a related conclusion: recipients may absorb a correction and then reinterpret its significance so that their prior attitude remains intact.⁶ The pertinent distinction is between belief accuracy and political evaluation. A responsible programme measures both, without assuming that the first necessarily governs the second.

Correction after exposure

Fact-checking has an indispensable role when a false claim is already circulating. It can identify the proposition, state what is wrong, supply an alternative explanation and point readers towards inspectable evidence. The continued-influence literature nevertheless shows that corrections may leave residues because the original account supplied a causal model which the correction removed without replacing.⁷ An effective correction should therefore be prompt, specific and explanatory. It should not merely stamp a label upon the content.

A two-wave experiment concerning a 2017 British election claim that EU immigration harmed the National Health Service illustrates both utility and limitation. A fact-check improved accuracy, but attribution to

³. Jens Hainmueller and Daniel J. Hopkins, ‘Public Attitudes toward Immigration,’ Annual Review of Political Science 17 (2014): 225–249.
⁴. Walter G. Stephan, Oscar Ybarra, and Kimberly Rios Morrison, ‘Intergroup Threat Theory,’ in Handbook of Prejudice, Stereotyping, and Discrimination, ed. Todd D. Nelson (New York: Psychology Press, 2009), 43–59.
⁵. Daniel J. Hopkins, John Sides, and Jack Citrin, ‘The Muted Consequences of Correct Information about Immigration,’ Journal of Politics 81, no. 1 (2019): 315–320.
⁶. Alvaro Hogue Rovelo, ‘Why Does Corrective Information Have a Muted Effect on Immigration Attitudes?’ American Politics Research (2025), <https://doi.org/10.1177/1532673X251324173>.
⁷. Stephan Lewandowsky et al., ‘Misinformation and Its Correction: Continued Influence and Successful Debiasing,’ Psychological Science in the Public Interest 13, no. 3 (2012): 106–131.

ideologically congenial or independent outlets affected reception, and the broader political significance of the correction remained bounded.⁸ This matters for migration policy because the messenger is never incidental. When institutions have become parties to the dispute, the authority printed beneath a correction may shape its reception as much as the evidence printed above it.

Prebunking before exposure

Prebunking derives from inoculation theory. McGuire and Papageorgis proposed that forewarning and exposure to weakened counterarguments could stimulate resistance to later persuasion⁹. Contemporary interventions adapt this principle by teaching users to recognise impersonation, scapegoating, emotional manipulation, false dichotomies and decontextualised evidence. The object is not to provide a catalogue of approved conclusions. It is to expose a modus operandi.

The Bad News game offers one of the better-known demonstrations. Across German, Greek, Polish and Swedish versions, 5,061 voluntary participants rated manipulative social-media posts as less reliable after the intervention; demographic variables did not substantially alter the reported inoculation effect.¹⁰ Earlier work on the English game likewise found improved resistance to common manipulation techniques.¹¹ Yet these studies primarily assess discernment or reliability judgements. They do not establish that a participant will maintain the gain for months, refrain from sharing politically congruent content in ordinary life, or become more willing to live and work alongside migrants.

A large Central and Eastern European campaign addressed narratives blaming Ukrainian refugees for economic hardship and social disorder. Six short videos were deployed in Poland, Czechia and Slovakia and reportedly achieved tens of millions of views. Post-campaign surveys showed improved recognition of manipulation in some settings, while performance varied by country, with localisation identified as a probable constraint.¹² This variation is not a footnote to success. It is evidence that translation is not localisation, reach is not exposure, and exposure is not durable behaviour.

Trusted messengers and the problem of borrowed legitimacy

A trusted-community-messenger approach begins from the reasonable observation that people evaluate sources as well as claims. Teachers, family doctors, local journalists, neighbourhood associations, employers and religious leaders may possess forms of proximity that a central ministry lacks. An evidence-based framework prepared by UNSW recommends familiar and relatable messengers, targeted channels, repetition and rapid response in communication about refugees and migrants.¹³ Familiarity, however, is not equivalent to legitimacy for every member of a community. A religious leader may be influential within one constituency and distrusted by another; a mayor may be recognisable yet politically polarising.

The messenger should therefore be treated as a testable component of the intervention. The pilot proposed below distinguishes the effect of message content from source identity and records baseline trust in each

⁹. William J. McGuire and Demetrios Papageorgis, ‘The Relative Efficacy of Various Types of Prior Belief-Defense in Producing Immunity against Persuasion,’ Journal of Abnormal and Social Psychology 62, no. 2 (1961): 327–337.
¹⁰. Jon Roozenbeek, Sander van der Linden, and Thomas Nygren, ‘Prebunking Interventions Based on the Psychological Theory of Inoculation Can Reduce Susceptibility to Misinformation across Cultures,’ HKS Misinformation Review 1, no. 2 (2020), <https://doi.org/10.37016/mr-2020-008>.
¹¹. Jon Roozenbeek and Sander van der Linden, ‘Fake News Game Confers Psychological Resistance against Online Misinformation,’ Palgrave Communications 5 (2019): 65.
¹². Jigsaw, ‘Defanging Disinformation Campaigns in Central and Eastern Europe,’ Google Safety Centre, 2023; see also Associated Press, ‘Google to Expand Misinformation Prebunking in Europe,’ 13 February 2023.
¹³. Susan Wodak et al., Countering Misinformation about Refugees and Migrants: An Evidence-Based Framework (Sydney: UNSW Kaldor Centre, 2025), 6–7.

candidate messenger. Taking this into account, the phrase ‘trusted messenger’ should be reserved for an empirically identified relationship rather than assigned ad hoc by programme staff.

Table 1 Comparative logic of the principal communication interventions

Intervention	Immediate strength	Characteristic weakness	Best initial use
Fact-checking	Corrects a specific circulating claim and creates a public record.	Arrives after exposure; accuracy gains may not alter attitudes.	High-reach false claim with accessible evidence and a credible alternative account.
Technique-based prebunking	Transfers recognition across claims that use the same tactic.	May be too general; durability and attitude effects are uncertain.	Recurring manipulation tactics before an anticipated information event.
Issue-based prebunking	Warns about a foreseeable narrative and supplies context in advance.	Risks repeating or advertising a claim that might not otherwise spread.	Predictable claims around policy changes, arrivals or elections.
Trusted messenger	May overcome institutional distance and adapt language locally.	Trust varies across groups and can be borrowed or exhausted.	Audiences with low central-institution trust and established local networks.
Combined approach	Joins preventive literacy to source credibility.	More costly and difficult to attribute effects.	Pilot settings where both narrative risk and institutional distrust are high.

Part II Comparative evidence audit

An audit standard for contested claims

A claim should not be designated misinformation merely because it is uncomfortable or because a public body disputes it. The evidence audit begins by fixing the proposition, population, place and period. ‘*Migrants increase crime*’ is analytically unusable until migrant, crime and increase have been defined. Does migrant mean foreign-born resident, foreign national, asylum applicant or irregular entrant? Does crime mean victimisation, police-recorded offences, suspects, charges or convictions? Is the comparison national, municipal or neighbourhood-level? Only after these questions are settled can the data carry argumentative weight.

Table 2 Evidence audit protocol

Audit stage	Question	Minimum evidence
1 Claim specification	What, precisely, is asserted, about whom, where and during which period?	Verbatim claim and a definition of all material categories.
2 Official position	What has the responsible authority stated?	Named statement, date and institutional remit.
3 Administrative trend	What changed in the recorded series?	Multi-year official data, numerator, denominator and breaks in series.
4 Analytical adjustment	Do age, sex, income, place or enforcement affect the association?	Peer-reviewed or transparent adjusted analysis.
5 Perception comparison	What does the public believe and how confident is the estimate?	Representative survey with wording and fieldwork dates.
6 Evidential verdict	How far does the proposition survive the audit?	Unsupported, misleading, partly supported, context-dependent, supported or inconclusive.

Sweden Crime migration and the denominator problem

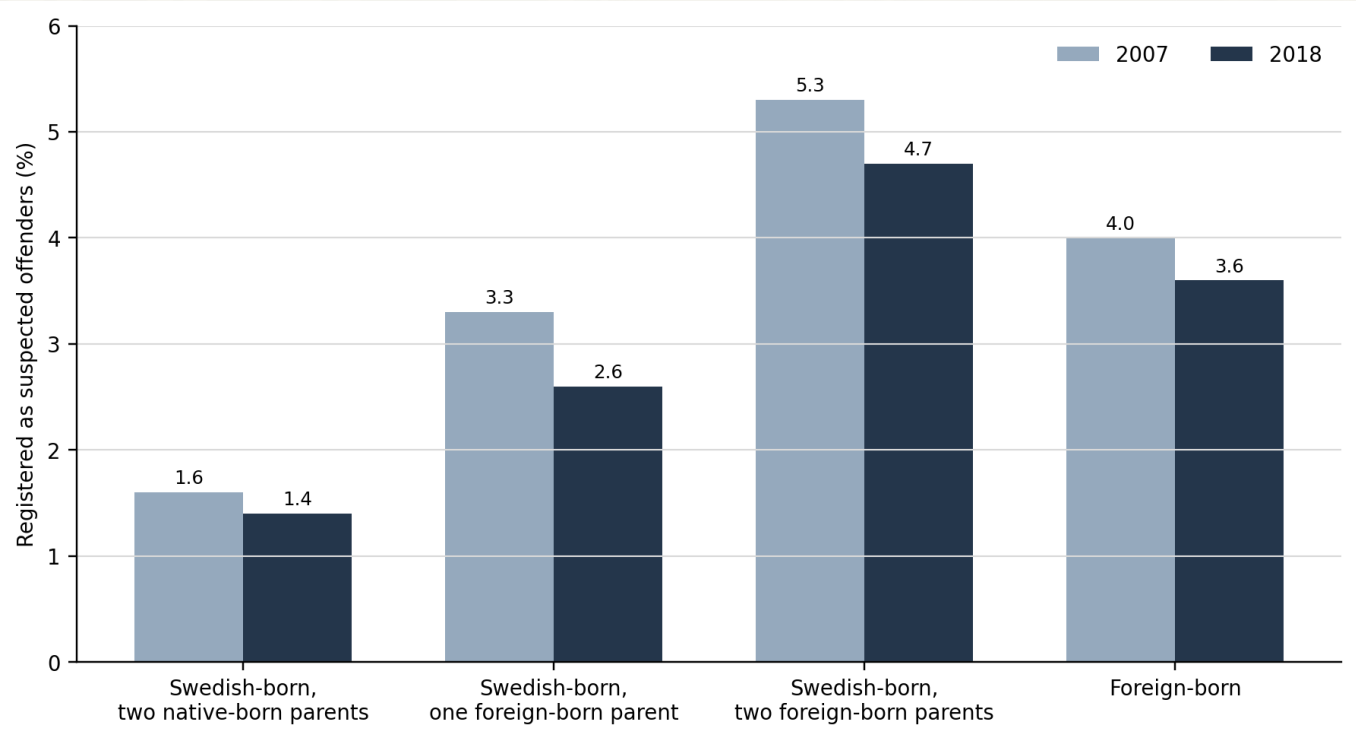
Sweden is useful precisely because the official evidence refuses an easy slogan. Brå’s 2021 study joined suspected-offender records with Statistics Sweden classifications and examined residents aged fifteen or over. For 2015–2018, 3.2 per cent of Swedish-born persons with two native-born parents were registered as suspected offenders, compared with 8.0 per cent of foreign-born persons and 10.2 per cent of Swedish-born persons with two foreign-born parents.¹⁴ It would be indefensible to erase these differences. It would be equally indefensible to present them without adjustment or trend.

When Brå accounted for age, sex, disposable income, education and municipality type, the foreign-born relative risk fell from 2.5 to 1.8; the relative risk for Swedish-born persons with two foreign-born parents fell

¹⁴. Swedish National Council for Crime Prevention, Registered Offending among Persons of Native and Non-Native Background: English Summary of Report 2021:9 (Stockholm: Brå, 2021), 5.

from slightly above 3 to 1.7.¹⁵ The adjustment did not eliminate the gap, and the report did not claim that it had. The result instead demonstrates how much explanatory work is concealed when a raw ratio is described as an inherent group propensity.

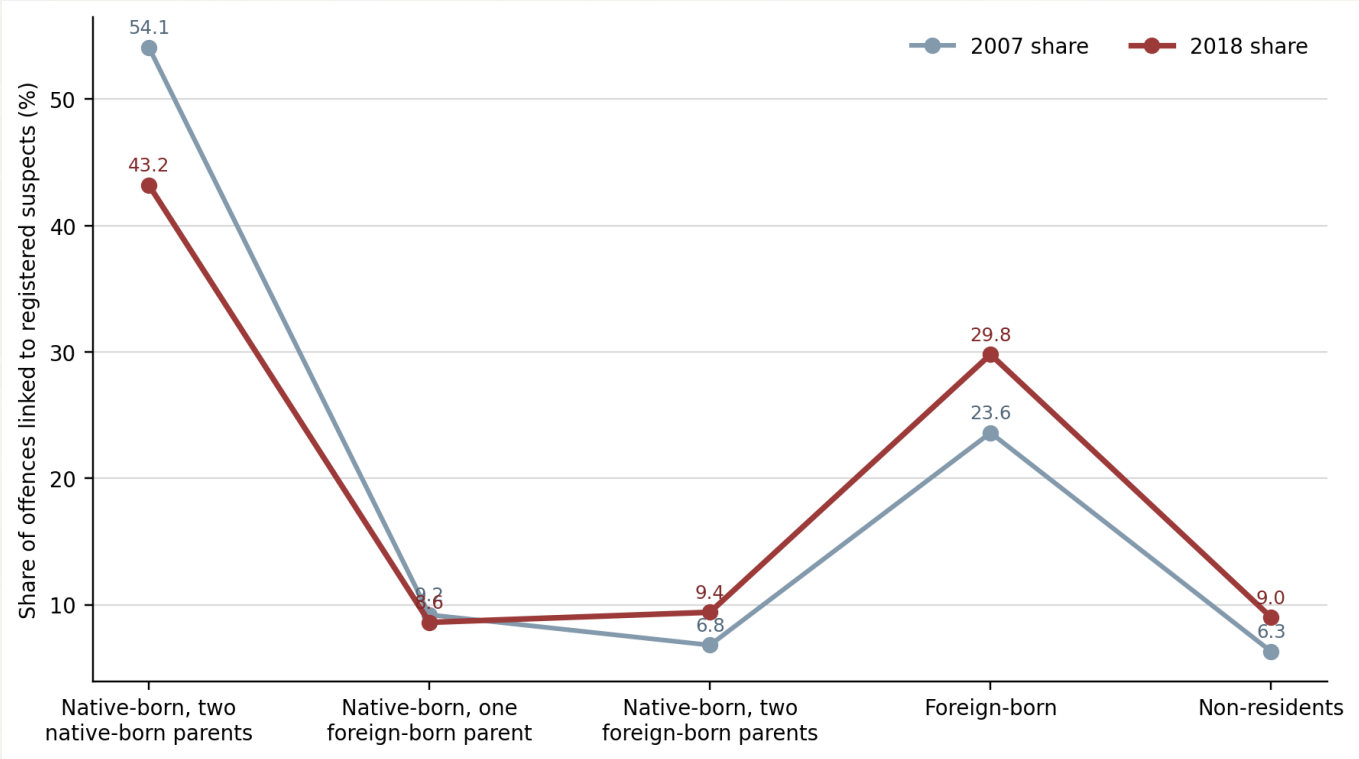
Figure 1 Share registered as suspected offenders in Sweden 2007 and 2018 Source Brå Report 2021 9



The time series produces a further complication. Between 2007 and 2018, the annual proportion registered as suspected offenders declined from 1.6 to 1.4 per cent among Swedish-born residents with two native-born parents, from 3.3 to 2.6 per cent among those with one foreign-born parent, from 5.3 to 4.7 per cent among those with two foreign-born parents, and from 4.0 to 3.6 per cent among foreign-born residents.¹⁶ Every group remained differently situated, yet the within-group direction was downward. A message asserting simply that ‘immigrant crime rose’ could point to an increasing number or share of offences linked to foreign-born suspects and still conceal the falling rate within a growing population.

16. Brå, Registered Offending, 8.

Figure 2 Distribution of offences linked to registered suspects in Sweden 2007 and 2018 Source Brå Report 2021 9



The distributional series shows that the share of registered offences linked to foreign-born suspects increased from 23.6 to 29.8 per cent, while the share linked to Swedish-born suspects with two native-born parents fell from 54.1 to 43.2 per cent.¹⁷ Brå attributed the broader shift primarily to changes in population composition. It also recorded offence-specific exceptions: assault linked to foreign-born suspects rose as a share, while the proportion of several other offence categories did not follow a general upward trend.¹⁸ The proper verdict is consequently context-dependent. The sweeping claim that immigration mechanically caused a national crime increase is not supported by this record; the claim that no group differences or offence-specific concerns existed is also contradicted by it.

17. Brå, Registered Offending, 11–12.

18. Brå, Registered Offending, 13–14.

Germany Official reporting and category discipline

Germany’s Federal Criminal Police Office has published a dedicated annual situation report on crime in the context of immigration since the reporting year 2015.¹⁹ Its category of Zuwanderer is not interchangeable with all foreign nationals or all foreign-born residents; depending on the report, it includes specified protection and residence statuses. Police Crime Statistics also register suspects rather than judicial findings and are influenced by offences that non-citizens may commit by virtue of migration law. These definitional boundaries belong in the main text, not in an apologetic appendix.

The German case accordingly serves a different function in this paper. It shows why an official chart cannot be copied into a campaign without reproducing its administrative ontology. A comparison between a status-defined migrant group and the entire German population would be particularly vulnerable to age and sex composition, exposure, residence duration and offence-mix effects. The audit verdict for a general causal assertion should remain inconclusive unless the claim is matched to a defined series and supplemented by adjusted research. Official publication improves transparency; it does not remove the analyst’s obligation to compare like with like.

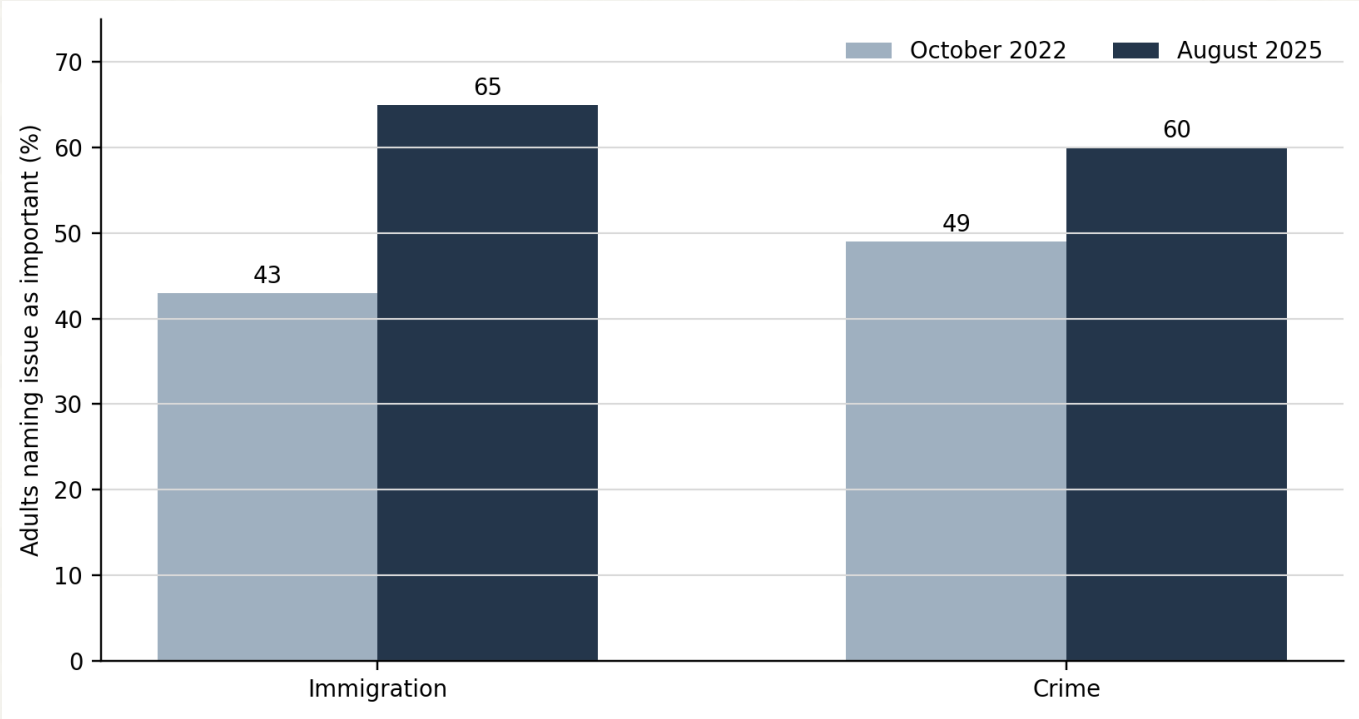
United Kingdom Data absence is itself an audit finding

The British case exposes a common institutional error: answering a question that the data do not answer. The Office for National Statistics has repeatedly explained that its principal crime series measure victimisation and police-recorded offences, and that it does not hold a comprehensive perpetrator breakdown by immigration or asylum status.²⁰ Prison nationality figures, visa enforcement records and Census country-of-birth estimates cannot be combined into a valid offender rate without linked populations, consistent definitions and exposure periods.

In such a setting, the honest verdict on a sweeping nationality claim may be inconclusive rather than false. This conclusion is professionally more useful than improvisation. It tells communicators not to publish an overconfident rebuttal and tells statistical authorities which linkage or publication gap is being exploited. It also separates the salience of immigration and crime from evidence that one causes the other. In Great Britain, the share of adults naming immigration as an important national issue rose from 43 per cent in October 2022 to 65 per cent in August 2025, while the corresponding figures for crime rose from 49 to 60 per cent.²¹ These parallel changes describe public concern; they do not establish a causal connection between the two issues.

19. Bundeskriminalamt, Kriminalität im Kontext von Zuwanderung: Bundeslagebild 2023 (Wiesbaden: BKA, 2024), methodological introduction.
20. Office for National Statistics, ‘Crimes in the UK by Race and Immigration Status,’ Freedom of Information response, 31 October 2024; Office for National Statistics, ‘Crime and Migration Statistics,’ Freedom of Information response, 4 February 2025.
21. Office for National Statistics, Public Opinions and Social Trends, Great Britain: August 2025 (London: ONS, 19 September 2025).

Figure 3 Public salience of immigration and crime in Great Britain Source ONS August 2025



Employment welfare and service pressure

The same discipline applies beyond crime. ‘Migrants take jobs’ collapses employment levels, displacement, wage effects, vacancy filling and sectoral dependence into one sentence. Labour-market effects may differ by skill, locality and time horizon, and an aggregate employment expansion may coexist with losses for particular workers. The migration literature generally finds modest average effects in receiving economies, but averages do not erase distribution.²² A defensible correction must state the level at which its conclusion holds.

Welfare claims present a parallel denominator trap. The number of foreign-national recipients does not disclose their share of total expenditure, their contribution through taxation, differences in household size, eligibility rules or refugee-specific transfers financed outside the ordinary benefit system. Where data disclose higher receipt in a recently arrived population, the finding should not be hidden. It should be explained with status, employment, age, household and programme rules. A prebunk should teach the reader

22. Christian Dustmann, Tommaso Frattini, and Ian P. Preston, ‘The Effect of Immigration along the Distribution of Wages,’ Review of Economic Studies 80, no. 1 (2013): 145–173; National Academies of Sciences, Engineering, and Medicine, The Economic and Fiscal Consequences of Immigration (Washington, DC: National Academies Press, 2017).

to ask ‘which benefit, which denominator and which population?’ before a viral total is allowed to stand for the welfare state as a whole.

Public-service pressure is still less amenable to a universal verdict. A municipality receiving a rapid inflow may experience genuine school, housing or primary-care pressure even where national fiscal effects are neutral or positive. Calling every such account misinformation would convert communication into denial. The intervention decision matrix below therefore contains an exit: when the audited claim rests upon a demonstrable service failure, the principal response is operational and distributive. Communication may explain the remedy, but it cannot be the remedy.

Central and Eastern Europe Prebunking refugee scapegoating at scale

The 2022 campaign concerning Ukrainian refugees in Poland, Czechia and Slovakia is the closest available operational precedent for the intervention proposed in this paper. Its importance lies neither in the novelty of short video nor in the size of the media purchase. The campaign attempted to intercept two recognisable narrative techniques before they consolidated: scapegoating refugees for economic hardship and using emotionally charged material to assign collective blame. The six videos did not ask viewers to adopt a position on admission policy. They demonstrated how an actor could move from a genuine anxiety over prices, employment or public provision to an unsupported attribution of responsibility. That distinction made it possible to address manipulation without requiring agreement on the whole politics of asylum.

The campaign reportedly reached roughly 38 million views across major platforms. Such a figure establishes distribution, not persuasion. Post-exposure research indicated gains in the ability to identify manipulative techniques and some reduction in willingness to share, but effects varied across the three national settings.²³ The weaker Slovak result is analytically valuable for the reason that it directs attention towards local political language, media habits, translation and source credibility. A communication asset can remain semantically accurate while failing to sound indigenous to the information environment in which it appears.

There is a further limitation. Technique recognition was measured more convincingly than durable social cohesion. The campaign did not establish that those who recognised scapegoating later revised their estimates of refugee benefit use, accepted a refugee neighbour or trusted public institutions more readily. The case therefore supports the scalability of technique-based prebunking, yet it also shows why a policy institution should resist reporting reach or immediate discernment as evidence that prejudice has receded. The proposed pilot retains the campaign’s preventive logic but adds delayed measurement of threat, trust and social distance.

23. Jigsaw, ‘Defanging Disinformation Campaigns in Central and Eastern Europe,’ Google Safety Centre, 2023; Associated Press, ‘Google to Expand Misinformation Prebunking in Europe,’ 13 February 2023.

Australia From rebuttal to an evidence based engagement framework

Australia offers a different type of case. The UNSW Kaldor Centre’s evidence-based framework assembles research on misinformation about refugees and migrants and converts it into practical principles: move quickly, repeat consistently, choose targeted channels, use familiar messengers and speak to the audience’s values and emotions.²⁴ The framework is valuable because it does not treat a correction as a self-executing package. It asks whether people will encounter the material, attend to it and regard the communicator as entitled to speak.

At the same time, a framework is not an impact evaluation. Its recommendations draw from a broader evidence base and supply a professional design standard; they do not demonstrate that every combination will work for every migration narrative in Australia or elsewhere. This distinction should remain visible when the model is transferred. Familiarity can increase attention, although a familiar messenger may reproduce local hierarchy or be distrusted outside a narrow constituency. Emotional language can make an explanation memorable, although it may also heighten the very threat whose influence the intervention seeks to reduce. The Australian contribution is thus best treated as a disciplined design repertoire which requires experimental testing, not as a catalogue of universal effects.

What the comparative cases establish

Read together, the cases disclose four different policy situations. Sweden has unusually detailed linked administrative analysis which permits a qualified verdict: meaningful differences exist, much of the raw gap is associated with demographic and socio-economic composition, within-group rates declined, and offence-specific exceptions remain. Germany publishes an official immigration-related crime series, but its administrative category cannot be converted into a rate for all migrants without careful denominator and status work. The United Kingdom illustrates a data gap in which confident perpetrator-by-status claims often exceed the national statistical architecture. Central and Eastern Europe supplies an intervention precedent which improved technique recognition at scale but produced heterogeneous national results. No single slogan can accurately summarise all four.

This comparison changes the order of institutional work. An agency should not commission a video, identify a spokesperson and only afterwards seek statistics to support the script. It should first determine what kind of evidential situation it faces. Where the claim is contradicted by an adequate series, a correction may be direct. Where a claim mixes accurate numbers with an invalid denominator, the communication task is explanatory. Where official data cannot answer the question, the institution should acknowledge the gap and avoid filling it with advocacy. Where a genuine local pressure exists, operational action takes precedence. Prebunking enters after this classification, because it is the manipulation technique rather than an administratively convenient conclusion that citizens are being taught to recognise.

24. Wodak et al., Countering Misinformation about Refugees and Migrants, 33–38.

The cases also expose a recurrent asymmetry. Public actors are often expected to prove the absence of a broad causal proposition, while purveyors of misinformation need only circulate a vivid incident. Evidence cannot abolish this asymmetry, but a stable audit protocol can prevent institutions from meeting exaggeration with counter-exaggeration. Over time, the consistency of the method may itself become a source of trust: the institution sometimes confirms a problem, sometimes qualifies it and sometimes refuses to pronounce where the data are insufficient.

Table 3 Migration narrative evidence and response matrix

Narrative	Typical manipulation	Evidence required	Possible verdict	Appropriate response
Crime	A salient offence is generalised to a population; suspects are presented as convicted offenders.	Multi-year rates, population denominator, offence type, status definition and adjusted analysis.	Often misleading or context-dependent; sometimes inconclusive.	Correction plus explanation; technique prebunking before predictable spikes.
Employment	Job creation and displacement are collapsed into a zero-sum national total.	Employment, unemployment, vacancies, wages, sectors, locality and time horizon.	Usually context-dependent.	Segmented evidence and policy response where losses are concentrated.
Welfare	Recipient counts are described as shares of all spending or as preferential entitlement.	Eligibility, recipients, expenditure, household structure, tax contributions and funding stream.	Claim-specific; frequently misleading.	Denominator prebunk and transparent budget table.
Public services	A real capacity failure is attributed solely to migration.	Demand, staffing, funding, ageing, regional allocation and service standards.	May be partly supported locally.	Operational remedy first; communication explains causes and action.

Comparative implications beyond the audited cases

The comparative record warns against treating national statistics as interchangeable units in a single rhetorical contest. Sweden can link population background to registered suspicion through administrative registers; Germany publishes a security category concerned with immigration-related status; the United Kingdom generally cannot answer the same question from routinely published perpetrator data. These are not merely differences in statistical maturity. They reflect distinct legal traditions, data-protection settlements, policing practices and political histories. A communication team that copies a Swedish denominator into a British claim, or treats the German category *Zuwanderer* as a synonym for all foreign-born residents, does not clarify the controversy. It manufactures a new category error while purporting to correct an old one.

The same discipline applies to temporal comparison. A rise in the number of recorded offences may coincide with population growth, altered recording rules, increased reporting by victims, a change in police priorities or a genuine deterioration in public safety. An honest audit therefore displays counts and rates together, specifies the denominator, identifies breaks in the series and distinguishes a change in composition from a change in individual propensity. Considering the fact that migration debates move rapidly from a single incident to a national generalisation, these qualifications cannot be left in technical annexes. They belong in the public explanation itself.

Local variation is equally consequential. National averages can conceal neighbourhoods in which school places, rental housing or primary care have come under unusual pressure. A national rebuttal may be arithmetically correct and experientially implausible to residents of such a locality. For that reason, the audit should be repeated at the smallest defensible geographical level before an intervention is commissioned. Where local data are unavailable, the institution should say so. The absence of evidence is not permission to convert a national mean into a description of every municipality.

Public-service narratives require a particularly careful division between demand and capacity. Migration can contribute to demand while still being an insufficient explanation of failure. Staffing shortages, capital investment, allocation formulae, ageing infrastructure and pre-existing regional inequality may carry greater explanatory weight. The appropriate response is consequently twofold: publish the contribution that population change plausibly made, and disclose the administrative decisions that shaped capacity. A communication campaign which names only the first factor becomes defensive; one which suppresses it becomes untrustworthy.

Employment claims present a related aggregation problem. Labour-market effects are distributed across sectors, occupations, places and time. A national estimate near zero does not prove that no incumbent worker experienced competition, just as a local wage effect does not establish an economy-wide loss. The public-facing evidence should therefore separate total employment, vacancies, wage distribution,

occupational segmentation and fiscal contribution. This does not weaken the response to misinformation. It prevents a defensible average from being used to erase a concentrated cost.

The implication for prebunking is important. Generic media-literacy content can teach recognition of scapegoating or false causality, but issue-specific material must be anchored in the categories actually used by the relevant state. The warning should show citizens how an administrative statistic can be stretched beyond its remit: suspects presented as offenders, recipient numbers presented as budget shares, foreign nationality presented as migration history, or a national total presented as a local rate. This form of prebunking does not instruct the audience what political conclusion to reach. It equips them to examine the inferential bridge between record and claim.

Part III A comparative behavioural pilot

Purpose and theory of change

The pilot is designed to answer a narrower and more useful question than whether prebunking works in the abstract. It asks which intervention, delivered by which messenger, improves discernment and reduces harmful sharing for a defined migration narrative, and whether any such gain changes perceived threat, trust or social distance over time. The programme should be piloted in two municipalities with contrasting levels of migration exposure and institutional trust. A sample of approximately 1,500 adults would permit five conditions of roughly 300 participants, subject to power calculations based on the selected primary outcome.

The theory of change is conditional. If participants receive a concise explanation of the manipulation technique before exposure, and if the explanation is delivered through a source they regard as legitimate, then they should recognise the tactic more accurately and hesitate before sharing. If the intervention also supplies a credible alternative account and does not dismiss genuine material concerns, the gain may reduce perceived threat. The second transition is less certain than the first and should be treated as an empirical question.

Table 4 Five arm pilot design

Arm	Intervention	Delivery	Primary comparison
A	Fact-checking of a circulating claim	Neutral visual card linked to inspectable official data	Post-exposure correction versus control.
B	Technique-based prebunking	Short video on scapegoating, false causality and missing denominators	Preventive literacy versus control.
C	Trusted-messenger explanation	Same factual content delivered by a pre-tested local messenger	Source effect versus Arm A.
D	Prebunking plus trusted messenger	Arm B content delivered by the same local messenger	Combined effect versus Arms B and C.
E	Attention control	Equal-length civic information unrelated to migration	Baseline for all treatment effects.

Intervention content and sequencing

Content development begins with the evidence audit rather than creative production. Researchers identify one high-reach claim in each selected municipality, fix its wording and classify its evidential status. Advisory groups containing statisticians, local service managers, migrant representatives, community organisations and political-diversity reviewers examine the draft. This process is not intended to negotiate the facts. It is intended to identify category errors, omitted local knowledge and phrasing that would inadvertently stigmatise a population or belittle a legitimate concern.

The prebunk should name the technique without repeatedly rehearsing the most inflammatory allegation. For example: ‘A single offence can be used to make a whole population appear dangerous. Before sharing, check whether the source reports an individual case or a population rate, whether it distinguishes suspicion from conviction, and whether the rate changed after population growth is taken into account.’ An issue-specific version may then use an innocuous or already familiar example. The fact-check, by contrast, should state the claim once, lead with the correction, supply the missing denominator and give an alternative causal account where evidence permits.

Messenger selection occurs before randomisation. A short baseline survey asks respondents to rate competence, honesty, independence and social proximity for candidate messengers. The programme may discover that no single actor commands broad confidence. In that event, it should stratify or use paired messengers rather than declare a winner by committee. The practical criterion is not celebrity but credible reach within the intended audience.

Measurement

Table 5 Outcome and measurement framework

Domain	Operational measure	Timing	Interpretive caution	
Belief accuracy	Agreement with verified and false claims using balanced item wording.	T0 T1 T2 T3	Accuracy may rise without attitude change.	
Discernment	Difference between ratings of credible and manipulative posts.	T0 T1 T2 T3	Avoid teaching to identical test items.	
Sharing intention	Likelihood of forwarding each item and an optional behavioural share task.	T1 T2 T3	Stated intention may overestimate restraint.	
Perceived threat	Perceived threat	Economic cultural and security threat subscales.	T0 T1 T2 T3	Do not combine distinct threat domains prematurely.
Trust	Institutional and messenger-specific trust.	T0 T1 T2 T3	Treatment may alter trust independently of accuracy.	
Social distance	Willingness for workplace neighbourhood school and family contact.	T0 T2 T3	Sensitive to social desirability.	
Durability	Treatment effect at four weeks and three months.	T2 T3	Attrition must be reported by arm.	

T0 is collected before intervention, T1 immediately afterwards, T2 after four weeks and T3 after three months. The primary endpoint should be preregistered, preferably discernment or willingness to share at T2. Multiple secondary outcomes require correction for multiple comparisons. The analysis should estimate intention-to-treat effects, report attrition and test whether baseline institutional trust, political identification, age, education and direct contact with migrants moderate results. Any subgroup analysis not preregistered should be labelled exploratory.

The programme should also record possible adverse effects. Repetition may increase familiarity; a government-branded prebunk may reduce trust among already sceptical participants; threat examples may heighten salience; and a local messenger may become exposed to harassment. A no-harm review should therefore examine not only mean effects but also whether any arm increases perceived threat, confidence in false claims or hostility among a defined subgroup.

Decision rules

Table 6 Pilot interpretation and programme decisions

Observed result	Interpretation	Decision
Accuracy and discernment improve; threat and social distance do not change.	The intervention builds informational competence but has not demonstrated social-cohesion effects.	Retain as media-literacy tool; do not advertise it as prejudice reduction.
Sharing declines at T2 and T3 without trust loss.	The intervention changes a relevant behavioural proxy and remains durable.	Scale cautiously with replication and platform-specific testing.
Trusted messenger outperforms identical institutional content.	Source legitimacy contributes independently to reception.	Invest in messenger networks while monitoring representativeness.
Prebunking effect disappears by T2.	The initial gain is short-lived.	Test booster exposure or abandon the format if cost-effectiveness is poor.
Threat increases in any arm.	The content may prime the harmful narrative or signal institutional alarm.	Pause deployment and revise content.
Audit finds a genuine service pressure.	Communication cannot remove the material cause of resentment.	Prioritise service capacity and publish the corrective action.

Governance ethics and transparency

A public prebunking programme operates close to the boundary between civic education and state persuasion. Safeguards must therefore be visible. The intervention should disclose its sponsor, publish the evidence audit, provide access to the source data, preregister the evaluation and permit independent replication. Political opinions should not be classified as misinformation merely because they are severe, unpopular or unfriendly to government. The object of correction is an empirically assessable proposition or a demonstrable technique of manipulation.

Migrant participation is necessary but must not reduce migrants to testimonial instruments. Advisory members should be paid, safeguarding procedures should cover exposure to hateful material, and examples should avoid repeating dehumanising language where the technique can be demonstrated otherwise. Personal data, political attitudes and social-distance responses require minimisation, secure storage and a clear retention schedule. Where the pilot uses platform advertising, targeting criteria and expenditure should be published.

Conclusion

Prebunking deserves a defined place within migration communication, although the term should not be allowed to promise more than the evidence can sustain. Its clearest function is preventive and epistemic. A well-designed intervention can teach citizens to recognise scapegoating, false causality, impersonation, emotional manipulation and missing denominators before one of these devices becomes attached to a politically charged incident. The strongest experimental findings concern discernment, reliability judgements and, in some studies, willingness to share. Evidence that the same intervention produces durable changes in perceived threat, institutional trust or willingness to accept social proximity is considerably thinner. The difference is decisive. Social cohesion cannot be inferred from a participant’s ability to identify a misleading post.

The matter is particularly delicate in migration because the disputed narrative often rests upon an observable event. A crime occurred; a waiting time increased; a benefit was paid; an employer hired a recently arrived worker. Misinformation enters through the relation constructed among these facts. An exceptional offence is converted into a population trait, a recipient count into a claim about the whole budget, or temporal coincidence into proof of causality. A preventive communication policy should make these inferential moves visible. It should also accept that some allegations will survive scrutiny in limited form. If a locality has experienced measurable service pressure or a specific group is overrepresented in a defined administrative record, denying that fact will not protect social cohesion. It will make the eventual correction less credible.

The Swedish case demonstrates what evidential honesty requires. Foreign-born residents and Swedish-born residents with two foreign-born parents were registered as suspected offenders at higher rates than the Swedish-born reference group during the period studied. Adjustment for age, sex, income, education and municipality type substantially reduced the excess risk but did not erase it. Over time, however, the annual proportion registered as a suspect declined in all four background groups, even while the distribution of offences linked to foreign-born suspects changed as the population changed. Each of these findings is true within the official record. Selecting only one produces advocacy; placing them together produces analysis.

Germany and the United Kingdom establish complementary cautions. Germany’s dedicated reporting improves visibility but relies upon a status category that cannot be treated as synonymous with migrants in general. The British statistical system, meanwhile, does not support many of the perpetrator-by-immigration-status comparisons demanded in public debate. In the former case, the danger lies in stretching an administrative category beyond its remit. In the latter, it lies in answering with unsuitable proxy data. An evidence-led institution must be prepared to state that a proposition is inconclusive when the necessary linkage does not exist. Such restraint may appear unsatisfying in a fast information cycle, but manufactured certainty is an expensive form of reassurance.

The Central and Eastern European campaign supplies a more encouraging lesson. Short preventive videos can be delivered at scale and can improve recognition of the techniques used to blame refugees for complex economic or social pressures. Yet the uneven national results also show that an intervention does not travel as freely as a media file. Language, humour, institutional reputation, political history and platform use affect reception. Localisation must therefore concern the social position of the message, not merely its translation. A trusted messenger may improve attention and credibility, but trust should be measured among the intended recipients rather than attributed by programme staff.

For these reasons, the proposed pilot separates four questions that public campaigns often collapse. Did the participant identify the manipulation? Did factual belief become more accurate? Did the participant become less willing to share the content? Did the intervention alter perceived threat, trust or social distance? An intervention may succeed on the first three and leave the fourth untouched. That is still a useful result, for the reason that reducing circulation has public value, but it should be reported as information resilience rather than prejudice reduction. Conversely, an intervention which increases accuracy while diminishing institutional trust may carry a cost that a single outcome measure would conceal.

The five-arm design provides a credible way to distinguish these effects. Fact-checking tests post-exposure correction; technique-based prebunking tests preventive recognition; a trusted-messenger arm isolates source effects; the combined arm tests whether prevention and local legitimacy reinforce one another; and an attention control provides a defensible baseline. Measurements at baseline, immediately after exposure, four weeks and three months prevent a short-lived reaction from being mistaken for durable change. Preregistration, intention-to-treat analysis, attrition reporting and a no-harm review are not technical

embellishments. They are the conditions under which a public institution may claim that its communication has worked.

The framework also establishes a boundary around communication. When an evidence audit identifies a real deficiency in housing, schooling, labour protection or public safety, the first intervention belongs to programme management and public administration. A municipality which lacks school capacity should not attempt to prebunk parents’ experience of overcrowding. It should expand capacity, explain the allocation decision and publish the timetable. Communication can prevent the failure from being falsely attributed to an entire population, but it cannot dissolve the failure itself. Taking this into account protects prebunking from becoming an ad hoc device for managing discontent that policy has helped to create.

A mature migration communication policy would therefore maintain three permanent capabilities. The first is an evidence unit able to audit claims rapidly and publish the source trail. The second is a network of locally tested messengers whose independence and representativeness are protected rather than merely borrowed. The third is an evaluation function capable of measuring delayed and unintended effects. These capabilities are more valuable than a library of campaign assets because they allow institutions to respond differently when the claim, audience or evidential condition changes.

The practical contribution of this paper is consequently a sequence of accountable decisions: specify the claim, establish the official record, test the denominator, review adjusted research, classify the evidential verdict, identify the manipulation technique, select and test the messenger, measure immediate and delayed outcomes, and publish both the result and its limitations. The order matters. Beginning with persuasion invites the evidence to become material for a predetermined message; beginning with the audit permits the message to remain answerable to reality.

Prebunking cannot remove political conflict over migration, nor should that be its purpose. Democratic disagreement includes legitimate disputes over admission, distribution, identity and public expenditure. The intervention becomes defensible when it equips citizens to recognise how evidence is being arranged and when it leaves them free to reach different policy judgements on an accurate basis. Its contribution to social cohesion is therefore neither unanimity nor institutional control of the narrative. It is a more difficult and more credible achievement: disagreement in which fabricated causality, collective blame and statistical sleight of hand possess less power than they did before.

Bibliography

Brå. Registered Offending among Persons of Native and Non-Native Background. English Summary of Report 2021:9. Stockholm: Swedish National Council for Crime Prevention, 2021.

Bundeskriminalamt. Kriminalität im Kontext von Zuwanderung: Bundeslagebild 2023. Wiesbaden: BKA, 2024.

Dustmann, Christian, Tommaso Frattini, and Ian P. Preston. ‘The Effect of Immigration along the Distribution of Wages.’ Review of Economic Studies 80, no. 1 (2013): 145–173.

European Commission, Joint Research Centre. What Are Europeans’ Views on Integration of Immigrants? Brussels: European Union, 2020.

European Parliament. The Impact of Disinformation Campaigns about Migrants and Minority Groups in the EU. EXPO/IDA/2021/653641. Brussels: European Parliament, 2021.

Hainmueller, Jens, and Daniel J. Hopkins. ‘Public Attitudes toward Immigration.’ Annual Review of Political Science 17 (2014): 225–249.

Hopkins, Daniel J., John Sides, and Jack Citrin. ‘The Muted Consequences of Correct Information about Immigration.’ Journal of Politics 81, no. 1 (2019): 315–320.

Horvath, Luca, et al. ‘Correcting Campaign Misinformation: Experimental Evidence from a Two-Wave Panel Study.’ HKS Misinformation Review 5 (2024).

International Organization for Migration. World Migration Report 2024. Geneva: IOM, 2024.

Jigsaw. ‘Defanging Disinformation Campaigns in Central and Eastern Europe.’ Google Safety Centre, 2023.

Lewandowsky, Stephan, Ullrich Ecker, Colleen Seifert, Norbert Schwarz, and John Cook. ‘Misinformation and Its Correction.’ Psychological Science in the Public Interest 13, no. 3 (2012): 106–131.

McGuire, William J., and Demetrios Papageorgis. ‘The Relative Efficacy of Various Types of Prior Belief-Defense in Producing Immunity against Persuasion.’ Journal of Abnormal and Social Psychology 62, no. 2 (1961): 327–337.

National Academies of Sciences, Engineering, and Medicine. The Economic and Fiscal Consequences of Immigration. Washington, DC: National Academies Press, 2017.

Office for National Statistics. Crime and Migration Statistics. Freedom of Information response. London: ONS, 2025.

Office for National Statistics. Public Opinions and Social Trends, Great Britain: August 2025. London: ONS, 2025.

Roozenbeek, Jon, and Sander van der Linden. ‘Fake News Game Confers Psychological Resistance against Online Misinformation.’ Palgrave Communications 5 (2019): 65.

Roozenbeek, Jon, Sander van der Linden, and Thomas Nygren. ‘Prebunking Interventions Based on the Psychological Theory of Inoculation Can Reduce Susceptibility to Misinformation across Cultures.’ HKS Misinformation Review 1, no. 2 (2020).

Stephan, Walter G., Oscar Ybarra, and Kimberly Rios Morrison. ‘Intergroup Threat Theory.’ In Handbook of Prejudice, Stereotyping, and Discrimination, edited by Todd D. Nelson, 43–59. New York: Psychology Press, 2009.

Wodak, Susan, et al. Countering Misinformation about Refugees and Migrants: An Evidence-Based Framework. Sydney: UNSW Kaldor Centre, 2025.

