Say No To Disinfo

Evidence-Based Practice Guide: Counter Polarisation Interventions

Executive Summary

Purpose

Polarization—particularly affective polarization—has increased in many democracies. High levels of polarization can undermine democratic norms, reduce willingness to cooperate across political differences, and increase support for undemocratic behaviour. A growing body of research evaluates interventions designed to reduce polarization. While most interventions produce modest effects, evidence suggests that carefully designed programs can reduce hostility and increase openness to political engagement across differences. This summary highlights practical interventions with the strongest empirical support and provides guidance on how they should be implemented.

Key Insights from the Research

1. Most interventions have small but meaningful effects

Experimental and field studies show that many depolarization interventions reduce affective polarization in the short term. However, effects are typically small, many effects decay after one to two weeks and outcomes are highly sensitive to intervention design. Interventions should therefore be viewed as incremental improvements, not single-step solutions.

2. Depolarization is easier than ideological agreement

Most successful interventions reduce hostility toward political opponents, not ideological disagreement. Reducing hostility can increase, willingness to engage with political opponents, support for democratic norms, openness to dialogue. Programs should focus on improving relationships between political groups, rather than eliminating political disagreement.

3. Design and delivery strongly affect outcomes

The success of interventions depends heavily on implementation. Programs are most effective when they ensure guaranteed exposure to participants, involve repeated engagement, are delivered through trusted institutions and combine multiple intervention types.

Say No To Disinfo

Evidence-Based Interventions

Research identifies three intervention families with the strongest evidence base.

1. Correcting Norm Misperceptions

- The problem: People frequently overestimate how hostile their own political group is toward political opponents. For example, partisans often believe members of their party are more supportive of political violence or exclusion than they actually are.
- The intervention: Providing accurate information about ingroup attitudes and norms can reduce hostility by correcting these misperceptions.
- Effective formats: visual “norm gap” graphics, short explainer videos, brief educational modules.
- Implementation principles: focus on participants’ own political group, present data visually, provide transparent sources, avoid moralizing language.

2. Humanizing Narratives

- The problem: Polarization is reinforced by stereotypes and abstract perceptions of political opponents.
- The intervention: Exposure to personal stories from individuals with different political views can reduce negative stereotypes and increase empathy.
- Effective narrative types: stories highlighting shared everyday experiences, narratives showing shared values but different political solutions, examples that challenge common stereotypes.
- Implementation formats: short video storytelling, podcast or audio narratives, serialized storytelling campaigns.
- Implementation principles: use first-person narratives, emphasize everyday experiences rather than ideological debate, avoid persuasion framing.

3. Structured Cross-Group Contact

- The problem: Limited interaction between political groups can reinforce mistrust and misunderstanding.
- The intervention: Structured dialogue and cooperative activities between individuals with different political views can reduce hostility. This approach builds on intergroup contact theory, which shows that interaction reduces prejudice when certain conditions are met.
- Evidence-based design conditions: facilitated discussions, politically diverse groups, equal participation, shared tasks or goals, multiple sessions over time.

4. Supporting Intervention: Improving the Information Environment

Polarization is also influenced by media and information ecosystems, including misinformation, selective exposure, and algorithmic amplification of conflict. Supporting interventions include, but are not limited to media literacy training, fact-checking partnerships, exposure to diverse viewpoints. These approaches strengthen the impact of other depolarization interventions.

Implementation Strategy

A common effective sequence is 1 - Norm correction to reduce exaggerated perceptions of hostility, 2 - Humanizing narratives to build empathy followed by 3 - Structured contact to create direct cross-group relationships. This progression moves participants from belief correction → perspective-taking → interaction.

Where to Implement Programs

Depolarization initiatives are most effective in environments that provide repeated interaction, institutional legitimacy, and guaranteed exposure to participants. Promising implementation settings include universities and educational institutions, workplaces and professional organizations, local governments and civic organizations and community-based programs.

Monitoring and Evaluation

Programs should track multiple indicators of polarization, including warmth toward opponents, perceived polarization, willingness to interact with opposing groups and support for democratic norms. Tracking both short-term and longer-term changes is essential to assess program effectiveness.

Key Takeaways

1. Evidence suggests polarization can be reduced, but effects are typically modest.
2. Interventions are most successful when they focus on reducing hostility rather than eliminating disagreement.
3. Program design is critical—poorly structured engagement can worsen polarization.
4. Repeated engagement and institutional support increase durability.
5. Combining multiple interventions produces the strongest results.

Say No To Disinfo

Evidence-Based Practice Guide: Counter Polarisation Interventions

Across the intervention literature reviewed, depolarization interventions generally produce small-to-modest improvements in affective polarization, typically measured as increased out-party warmth on feeling thermometers. The large meta-analytic synthesis reports an average improvement of ~5.4 points on a 101-point scale, and it emphasizes that these effects decay rapidly, with ~75% of the reduction decaying within about a week, and most regressing to baseline beyond ~2 weeks.

This means the “most promising” interventions are not those with the biggest immediate lift in a lab, but those that can be:

- delivered with guaranteed exposure,
- embedded in institutions / repeated contexts,
- paired with structural or information-environment supports
- prioritize careful design, testing, and adaptation to context

Depolarization interventions often produce small and context-dependent effects. Across the literature, outcomes vary substantially depending on design, context, and participant composition. Some interventions reduce affective polarization, while others show null or even polarizing effects when implemented poorly. This makes design quality and implementation conditions as important as the intervention type itself.

The Voelkel et al. megastudy summary shows that 23/25 treatments reduced partisan animosity immediately, but only a smaller subset showed durability at 2 weeks, and there was limited evidence of durable effects on anti-democratic attitudes. Therefore for a UK/EU strategy, success criterion should be durability and real-world implementation feasibility.

The megastudy synthesis explicitly flags that in real-world settings it’s hard to ensure people are exposed to treatments, and that in uncontrolled environments counterframes can erode effects. Therefore the best interventions are those that can be deployed through: institutions (universities, workplaces, local councils), media/platforms (where impressions can be guaranteed), and networks that can provide repeated and credible exposure.

The syntheses converge on three interventions as the most consistently efficacious for reducing partisan animosity / affective polarization:

1. Correcting misperceptions (especially norm perceptions)

The systematic reviews note that researchers have reduced negative partisan attitudes and even support for partisan violence by reducing misperceptions about the prevalence of negative

attitudes and support for violence. The same review also highlights that partisans often hold false beliefs about the outgroup's composition and characteristics, and correcting these misperceptions can reduce animosity.

Crucially, the intervention-focused review by You et al. suggests the most promising approach is to focus on partisans' perception of their ingroup's normative levels of outgroup bias i.e., ingroup norm perception correction rather than only outgroup stereotyping.

2. Humanization via sympathetic narratives / perspective-getting

The direct-meta-analysis notes that among the strongest interventions used in subsequent large-scale experiments were sympathetic personal narratives and appeals to a common identity.
The systematic review also highlights evidence that exposure to opponents' thoughtful arguments and personal experiences reduces negative impressions, and that dialogue training approaches often emphasize personal experience and receptive engagement.

3. Structured, positive intergroup contact

The systematic review emphasizes that contact can reduce partisan animosity, but that form matters: not all contact works, and some can backfire. It emphasizes classic conditions that make contact effective (equal status, shared goals, cooperation, authority support), and notes the scalability–durability trade-off (deep durable contact is harder to scale; scalable interventions often don't produce durable offline change).

The literature also highlights two consistent insights suggesting a complementary fourth intervention:

1. Media exposure can both increase and decrease polarization.
2. Media literacy and critical information skills help people identify polarizing or biased content.

4. Information Environment & Media Literacy

Polarization is partly driven by information dynamics including selective exposure, misinformation, algorithmic amplification and biased or uncivil political communication. Strengthening citizens' ability to interpret information critically can reduce the impact of polarizing content. The evidence shows that media exposure can both increase and decrease polarization depending on the content encountered, exposure to diverse information, fact-checking, and counter-attitudinal content can reduce ideological polarization in experiments.

Say No To Disinfo

Core Principles for Designing Depolarisation Interventions

Across the literature, several design principles consistently determine whether interventions succeed or fail.

1. Design matters as much as intervention type

Many interventions show mixed results across studies because outcomes depend heavily on design. For example, deliberation reduces polarization in many studies but increases it in others when poorly structured, and information provision frequently has no effect, and sometimes increases polarization. Programs should therefore treat interventions as design-sensitive tools, not universal solutions.

2. Target affective polarization first

Most successful interventions reduce affective polarization (hostility toward opponents), not ideological disagreement. Reducing hostility can increase willingness to engage, openness to dialogue and tolerance of democratic disagreement. Programs should not aim to eliminate political disagreement.

3. Repetition is essential

Many interventions produce short-term attitude change but effects decay rapidly. Durability requires repeated exposure, institutional embedding, follow-up engagement.

4. Institutional delivery increases success

Interventions are most effective when delivered through organizations that provide repeated interaction, legitimacy and trust, and guaranteed exposure. Examples include universities, workplaces, local authorities, civic organizations.

Say No To Disinfo

Intervention 1: Ingroup norm / misperception correction

Ingroup norm / misperception correction interventions reduce polarization by correcting exaggerated beliefs about how hostile *my own side* is toward opponents, and/or how extreme, violent-accepting, or anti-democratic *my own side* is.

The mechanism updating what people think is normal for their group, which then shifts their personal attitudes through assimilation. This is singled out in the review synthesis as a particularly promising approach: directly tackling perceptions of the ingroup's normative bias against the outgroup. From the systematic review and meta-analysis, there are four reasons this is a promising:

1. Misperceptions are widespread (about outgroups, and about what is normal in politics), and correcting them can reduce animosity.
2. Corrections can also affect high-stakes outcomes like support for partisan violence when targeted at relevant misperceptions.
3. Norm-based approaches can be highly scalable (short messages, digital delivery), which matters given the exposure constraint.
4. Norm-correction fits with "wise intervention" principles in the systematic review: it can satisfy motives for accuracy and belonging while avoiding reactance if done non-moralistically.

Some important norms for a UK/EU context are as below:

- Norms about violence and harassment (e.g., "Most supporters of your party reject political violence/harassment.")
- Norms about democratic procedures (accepting election loss; rejecting rule-breaking)
- Norms about everyday civility and willingness to talk/work across differences

Targeting rules

- Correct ingroup norms for the respondent's self-identified political ingroup (party, ideological camp, or salient identity grouping).
- Optionally, pair with outgroup metaperception correction as a comparator.

How to implement it?

Formats

Start with three “assets” you can deploy everywhere:

1) Norm-gap card (best baseline unit)

A single graphic that shows: “What you guessed people in your group believe/do”
Vs “What people in your group actually report”. Keep it to one norm per card (don’t bundle multiple corrections).

2) 60–90 second explainer

A short video where a credible messenger explains: the norm gap, why misperceptions occur (loud minorities, media incentives), the accurate norm.

3) 3–5 minute institutional microlearning

A small module embed-able in: university induction, workplace training, union/professional body learning and local authority civic programmes.

Include a brief reflection prompt (to reduce reactance and increase internalization).

Channels

1. Universities and further education colleges (high reach into young adults; repeatable cohorts)
2. Large employers / professional bodies (guaranteed exposure; social norm relevance)
3. Local authorities / community hubs (local trust; place identity)
4. Public-interest media (broad reach, legitimacy)
5. Paid digital distribution (for targeted reach when access is limited)

Evidence-based design rules (to prevent backfire)

These are the “musts” implied by the reviews’ backfire concerns and wise-intervention principles:

- Neutral tone (no scolding; avoid “you were wrong”)
- Transparent sourcing (who collected data; sample size; date)
- Avoid identity threat (“people like you are better than assumed,” not “your side is bad”)
- Minimise cognitive load (one correction at a time)
- Pretest comprehension and trust (brief panel tests)

What success looks like?

Polarization should ideally be measured across multiple dimensions:

- Affective polarization (outgroup warmth) / feeling thermometers (multiparty adaptation)
- Perceived norms (did the misperception move?) and Ideological distance
- Dehumanization and threat perception
- Sentinel outcomes (violence acceptance; democratic norm support)
- Willingness to engage with opponents.

Because durability is the bottleneck in the meta-analysis, measure at:

- immediate post
- 1 week
- 2+ weeks / 1 month

Intervention 1 Pack

Design principles

Principle 1: Ingroup first, always. Every asset must be framed as: "Most people in your political group…"
The systematic review flags ingroup norm perception as especially promising; psychologically, ingroup norms are assimilated rather than debated. Information that relates directly to recipients' interests or experiences has stronger effects.

Principle 2: One norm per asset

Each asset should correct one and only one belief: violence acceptance, harassment, democratic procedures and willingness to engage civilly. Bundling increases cognitive load and reactance; misperception correction works best when the "gap" is obvious.

Principle 3: Make the "norm gap" explicit

Each asset must visually or verbally show perceived norm ("what people think"), vs actual norm ("what people report"). Norm change requires contrast. Without a visible gap, people reinterpret the information defensively.

Principle 4: Radical transparency of source

Always include: who collected the data, when, sample size and exact wording (or a link). The meta-analysis stresses counterframes and distrust as real-world constraints; transparency reduces accusations of manipulation. Infographics and video are more effective than text for correcting misperceptions.

Principle 5: No moral judgement

Avoid "should" language, evaluative framing (“better”, “healthier”) or implied blame (“you thought wrongly”). Use descriptive language (“most supporters report…”). Norm correction fails when it becomes moral instruction; people resist being told their group is “bad”.

Pack components

Minimum viable pack

1. 6–10 norm-gap cards segmented by political ingroup (party or bloc)
2. 2 short explainers (60–90s) one on *what the norm is*, one on *why misperceptions arise*
3. Source & methods sheet
4. Measurement items perceived norm before/after

Gold standard pack (for institutions)

- Microlearning module (3–5 min)
- Reflection prompt (“Did this surprise you? Why?”)
- FAQ / scepticism handling (“How do we know this?”)

Delivery & iteration guidance

1. Institutional channels first (universities, employers, local authorities)
2. Public-interest media second
3. Paid digital distribution only where necessary

The synthesis identifies exposure as the binding constraint; institutions guarantee exposure.

Iteration signals

Iterate if perceived norms shift but hostility does not → pair with Intervention 2, trust is low → strengthen source credibility / messenger, effects decay fast → embed in repeated institutional touchpoints

Common failure modes

Targeting outgroup norms only, bundling multiple norms, vague sourcing (“research shows…”), moralising tone and deploying in low-trust channels first.

Say No To Disinfo

Say No To Disinfo

Intervention 2: Sympathetic narratives & perspective- getting

Across the meta-analyses and systematic reviews, narratives and perspective-getting repeatedly appear among the strongest and most reliable interventions for reducing partisan animosity, especially in large-N experiments.

Key findings from the synthesis:

- Sympathetic personal narratives and appeals to common identity were among the most effective treatments for reducing partisan animosity immediately
  .
- The exposure to opponents' thoughtful arguments and lived experiences reduces negative impressions and hostility, especially when participants are trained to listen receptively rather than debate
  .
- Dialogue-training and perspective-getting approaches are consistently associated with reductions in dehumanisation and moral contempt, even when policy disagreement remains.
- However, effects decay quickly unless narratives are repeated or embedded, narratives do not reliably shift democratic norm support on their own and poorly designed narratives can be dismissed as propaganda or trigger cynicism

Why is this intervention promising?

Narrative-based interventions work through a different psychological mechanism than norm correction:

- Norm correction updates beliefs about "what people like me are like."
- Narratives update who the other side is replacing abstract caricatures with individuated humans.

From the evidence base, narratives are especially strong at reducing dehumanisation, lowering moral absolutism ("they are evil") and increasing willingness to engage or listen,
even when they do not change policy preferences. This makes Intervention 2 particularly valuable in the UK + EU context because political divisions are often cross-cutting and identity-based (region, class, migration, culture), multiparty systems make "the outgroup" diffuse, and people may reject overt "corrections" but still engage with stories.

Narratives are best used as a mass-reach humanisation layer, and a recruitment funnel into deeper engagement (e.g. structured contact).

What kinds of narratives does the evidence support?

The research summarised does not support generic “we are all the same” messaging. Instead, the most effective narratives share specific features.

High-performing narrative archetypes

From the reviews and the dialogue literature, four archetypes consistently perform better:

1) “Relatable life, different politics”

The same town / region / job / family role, different party or political stance and emphasises everyday life, not ideology

2) “Shared value, different path”

Both sides care about the same moral value (safety, fairness, opportunity) but disagree on how to achieve it. This reduces moral condemnation while preserving disagreement

3) Cross-cutting identity narratives

- Individuals whose identities break stereotypes (e.g. union member voting conservative; migrant small business owner sceptical of immigration)
- Particularly effective in multiparty Europe

4) Moral humility stories

“I realised I was wrong about them”. This one is powerful but risky if preachy (see failure modes below)

How to implement Intervention 2 well?

Formats

1) Short videos (60–120 seconds). These are highest engagement, works well on platforms and broadcasters and should be first-person, not narrated commentary

2) Audio + portrait + transcript. Lower production cost, strong for local radio, podcasts, community channels and particularly effective in regional European contexts

3) Serial storytelling (critical). Because effects decay quickly, one-off stories are insufficient. Use weekly or fortnightly series, rotate protagonists across cleavages

4) Facilitated perspective-getting scripts. Used inside workshops, classrooms, or contact programmes. Participants listen to a story and reflect, rather than argue

Channels

1. Local and regional media (place identity is a strong bridge in Europe)
2. Universities and further education colleges (student comms, curricula, societies)
3. Employers and professional bodies (internal comms, learning platforms)
4. Community and civic networks (libraries, hubs, faith networks)
5. Paid digital distribution (when access is otherwise limited)

Design rules to prevent backfire

To stay evidence-aligned: Use first-person voice only. Avoid conversion narratives (“I saw the light”), moral instruction (“we should all…”) and elite or abstract language. Include ordinary life details (work, family, routines) and pretest for trust, perceived manipulation and reactance

Rule of thumb: If a story sounds like a campaign message, it will likely fail.

What success looks like?

Reductions in dehumanisation and moral contempt, increased out-party warmth and increased willingness to engage or listen.

Given decay concerns, measure: immediate post, 1-week, and ideally repeated exposure effects over time

Intervention 2 Pack

Design principles

Principle 1: First-person lived experience only.
Narratives must be in the subject’s voice. No expert commentary explaining “the lesson”.
Narratives work through empathy and transportation, not argument.

Principle 2: Political difference must be explicit but not central.
The political stance must be clear (“I vote X”, “I support Y”). The story should not revolve around persuading or justifying it. Hidden politics feels deceptive; overt persuasion triggers reactance.

Principle 3: Identity bridge is mandatory.
Each narrative must contain at least one bridge: shared place, shared role (parent, worker), and shared value (safety, dignity). Bridges allow empathy without agreement..

Principle 4: Ordinary detail > ideology.
Include daily routines, family moments, frustrations and trade-offs Avoid slogans, abstract principles and manifesto language. Ordinary detail signals authenticity and defeats caricature.
Principle 5: Serial exposure beats single stories.
Packs should be designed as series, not one-offs. The meta-analysis shows rapid decay; repetition with variation sustains effects.

Pack components

Minimum viable pack

1. 12–20 narratives
2. Tagging schema (cleavage, bridge, format, intensity)
3. Reflection prompts
4. Use guidance of when to deploy which story
5. Recruitment CTA link to structured contact (Intervention 3)

Gold standard pack

- Narrative commissioning brief
- Local trust-broker sign-off checklist
- Comment moderation guidance (for public release)

Delivery & iteration guidance

Pair narratives with: light norm correction (1), or invitations to deeper engagement (3).

Iteration signals

- If stories feel “nice but irrelevant” → strengthen political explicitness
- If cynicism increases → audit authenticity and messenger
- If engagement drops → shorten, localise, or re-sequence

Common failure modes

- elite-produced “stories about others”
- conversion narratives (“I changed sides”)
- moral instruction
- single hero story representing “the other side”

Say No To Disinfo

Say No To Disinfo

Intervention 3: Structured positive intergroup contact

Intergroup contact is one of the most studied and most condition-dependent depolarisation interventions. From the evidence:

- Contact can reduce partisan animosity, but only when structured correctly; some forms of contact show null or negative effects.
- The review reiterates classic conditions under which contact is effective:
    - equal status,
    - shared goals,
    - cooperation,
    - authority or institutional support.
- The review also stresses a scalability–durability trade-off:
    - deeper, repeated contact → more durable effects,
    - but harder to scale

The evidence explicitly warns that poorly prepared contact can amplify stereotypes and polarisation rather than reduce them. Despite these risks, structured contact is critical because:

1. It offers greater durability than one-shot messaging
2. It builds expectations of cooperative disagreement, not just attitudes
3. It can generate new ingroup norms (“people like me can talk across difference”)

In other words:

- Interventions 1 and 2 reduce hostility,
- Intervention 3 converts those reductions into social practice.

This is why the evidence suggests positive, semi-structured discussions via online platforms (e.g. Zoom) as a promising compromise between scale and durability

The two most evidence-aligned contact models (UK + EU)

Model 1: Cooperative contact cohorts (durability-oriented)

Structure

- 6–10 participants, mixed political identities
- Work together on a shared local problem
- 4–6 sessions over 4–8 weeks

Why it works

- Cooperation reduces threat

- Shared goals shift attention away from identity conflict
- Repetition builds trust

Best settings

- Universities
- Workplaces
- Local authority programmes

Model 2: Perspective-getting cohorts (safety-oriented)

Structure

- Structured “ask-and-listen” format
- No debate or persuasion
- Facilitated rounds focusing on lived experience and values

Why it works

- Lower risk of escalation
- Strong reductions in dehumanisation
- More acceptable in high-tension environments

Best settings

- Community organisations
- Faith and civic networks
- Early-stage engagement before cooperative work

Implementation conditions

- Trained facilitation (unmoderated contact is high risk)
- Equal speaking time
- Clear norms and enforcement
- Voluntary participation
- Institutional backing (visible authority support)
- Multi-session design (one-offs are weak)

Delivery formats that scale

- Hybrid / Zoom cohorts allow cross-region scaling while retaining structure
- Chapter or cell models allow local ownership with central standards
- Train-the-trainer pipelines reduce cost and increase diffusion

Failure modes to avoid

- Debate-style formats increase antagonism
- Single-session contact has weak effects
- Power asymmetries undermine trust
- Poor facilitation amplifies stereotypes

Success criteria

Appropriate outcomes include sustained reductions in affective polarisation, increased willingness to re-engage, formation of cross-political ties and reduced fear and threat perception.

Measurement should extend longer than for other interventions: immediate, 1 month and 3+ months where feasible.

Intervention 3 Pack

Design principles

Principle 1: Structure beats goodwill.
Contact must be designed, not left to goodwill. The systematic review warns that unstructured contact can backfire. Physical interaction reduces hostility more reliably than online formats.

Principle 2: Multi-session needed.
Minimum: 4 sessions. One-offs are insufficient and risk harm.

Principle 3: Cooperation before disagreement.
Start with: shared goals, joint problem-solving and values exploration. Delay issue disagreement.

Principle 4: Facilitation is paramount.
Facilitators must: enforce norms, manage power asymmetries and intervene early.

Principle 5: Equal status must be actively protected.
Watch for class, educational and linguistic dominance. Status imbalance nullifies contact benefits. Equal speaking time, Reason-giving, Respectful disagreement.

Principle 6: Heterogeneous groups.
Mixed political viewpoints outperform homogeneous groups.

Pack components

Minimum viable pack:

4–6 session curriculum, facilitator script, recruitment & screening guide, safety & escalation protocol and fidelity checklist

Gold standard pack:

train-the-trainer materials, session observation rubric and participant reflection tools

Delivery & iteration guidance

Best testbeds: Universities, employers/professional bodies and local authority community hubs

Iteration signals

- If tension spikes early → slow pace, strengthen norms
- If participants disengage → clarify goals and value
- If dominance appears → tighten facilitation rules

Common failure modes: Debate framing, single session, untrained facilitators, forced participation and lack of institutional backing.

Supporting Layer: Information Environment and Media Literacy

Polarization is partly driven by the information ecosystem, including selective exposure, misinformation, and algorithmic amplification. Evidence suggests that improving media literacy and exposing citizens to diverse viewpoints can mitigate these dynamics.

Implementation options

Digital literacy modules

Training that helps participants identify biased sources, misleading narratives and polarizing rhetoric.

Fact-checking partnerships

Collaboration with trusted journalists to provide credible corrections.

Diverse information exposure

Platforms and media outlets can promote exposure to counter-attitudinal content.

Complementarities between interventions

1 → 2: norm correction increases receptivity to narratives

The intervention-focused review suggests directly targeting ingroup norm perceptions about outgroup bias is a particularly promising way to reduce affective polarization. When people believe “my side is not as hostile as I assumed,” the identity-threat temperature drops, and narratives are less likely to be rejected as “propaganda” or “enemy apologism.”

Practical design: in a rollout, deliver a norm-correction asset first, then a narrative asset.

2 → 3: narratives are the best recruitment funnel into structured contact

The systematic review highlights that exposure to personal experiences, receptiveness, and dialogue skills can improve cross-partisan perceptions and make conversations more productive. Narratives soften threat perceptions and make people more willing to enter higher-touch settings like facilitated cohorts.

Practical design: every narrative pack should contain a “next step” invitation to join a cohort/contact activity.

3 → 1 (and 2): contact creates lived proof that reinforces new norms and humanization

The systematic review stresses that contact reduces animosity when structured properly, and that deeper engagement is more likely to yield durable effects. Once people experience respectful engagement, their sense of what’s normal (“people like me can talk across differences”) becomes experiential, reinforcing norm shifts and humanization.

Practical design: close each contact session with a short reflection prompt that makes norms and humanization explicit (“What surprised you about how ‘your side’ showed up?” / “What did you learn about them as people?”).

Recommended sequencing

Layer 1 – Norm correction
Reduce exaggerated beliefs about hostility.

Layer 2 – Humanising narratives
Replace stereotypes with human stories.

Layer 3 – Structured contact
Create repeated interaction across political difference.

Supporting layer – Information environment
Strengthen resilience to polarising media dynamics.

Each layer addresses a different driver of polarization.

Failure interactions (how interventions 1–3 can undermine each other)

A) 3 (contact) implemented poorly can undo gains from 1 and 2

The review warns that not all contact reduces animosity and that some forms can exacerbate it; the form and preparation matter. So if contact is unmoderated or debate-like, it can increase hostility and cynicism, making subsequent norm correction and narratives less credible.

1 (norm correction) with weak source credibility increases skepticism about 2 and 3

The meta-analysis synthesis highlights the real-world challenge of exposure and counterframes; if people distrust the message, it becomes counterproductive. Skepticism then generalizes to narratives (“staged”) and contact (“manipulation”).

2 (narratives) that feel preachy reduces willingness to join contact (3)

The systematic review emphasizes the motivational challenge and the need to avoid moralizing language in dialogue contexts. If narratives are moral instruction rather than lived experience, they trigger defensiveness.

Say No To Disinfo

Counter-Polarisation Evaluation Framework

Evaluation Architecture

Every intervention should be evaluated across four linked layers:

Layer	What you're measuring	Why it matters
Outcomes	Changes in polarisation	Core impact
Mechanisms	Why change happened	Improves intervention design
Behaviour	Real-world effects	Bridges lab → practice gap
Implementation	Delivery quality	Explains success/failure

Most studies measure outcomes only, but miss mechanisms and implementation, limiting real-world usefulness.

Core Outcome Measurement Framework

1. Primary Outcome: Affective Polarisation

Measurement tool: Feeling thermometer (0–100 scale)

Measure:

- Warmth toward ingroup
- Warmth toward outgroup
- Calculate:
    - Polarisation gap = ingroup – outgroup

Why this is critical: Most consistent measure across studies, Most responsive to interventions

2. Secondary Outcomes

1. Dehumanisation / social distance (“How evolved/civilised is the other group?” or Willingness to act for: neighbour, colleague, family member)

2. Perceived threat (Cultural/Economic/Political)

3. Democratic norms (Acceptance of opposition legitimacy, election outcomes)

4. Support for political violence

These are harder to shift but high-impact if changed.

Mechanism Measurement Framework

This must be tailored to the intervention type.

Intervention 1: Norm Correction

Measure:

- Perceived ingroup attitudes
- Perceived outgroup attitudes
- Actual attitudes (for gap calculation)

Derived metric:

- Misperception gap

Intervention 2: Narratives

Measure:

- Perceived intentions of outgroup
- Empathy
- Moral judgement (e.g. “good vs bad people”)
- Perceived common ground

Intervention 3: Contact

Measure:

- Trust

- Willingness for future interaction
- Anxiety before/after contact
- Quality of interaction

Intervention 4: Information environment

Measure:

- Trust in information sources
- Perceived credibility
- Exposure to polarising content

Behavioural Measurement Framework

A. Low-cost behavioural proxies

- Click-through on cross-group content
- Time spent engaging with opposing views
- Voluntary sign-up for dialogue

B. Medium-strength indicators

- Participation in:
    - community events
    - deliberative forums

C. Strong behavioural indicators

- Actual cross-group collaboration
- Workplace or school interaction patterns

Implementation & Fidelity Measurement

- Exposure (Did participants actually see the intervention?)
- Engagement (Completion rate/Attention/Comprehension)
- Trust (Did participants believe the content?)
- Fidelity (Was the intervention delivered as intended?)

Why this matters: Null effects may reflect poor delivery, not ineffective design.

Say No To Disinfo

Evaluation Design Options

A. Minimum viable design

- Pre–post survey
- Comparison group (if possible)

B. Recommended (pilot level)

- Randomised controlled trial (RCT)
- Two groups: treatment and control

C. Best practice

- Multi-arm RCT: e.g. different narrative types
- Allows mechanism testing and optimisation

Measurement Timeline

Minimum timeline

- T0: Baseline
- T1: Immediate post
- T2: 1 week
- T3: 2–4 weeks

Gold standard

    - T4: 2–3 months

What to analyse

- Immediate impact
- Rate of decay
- Residual effect

Segmentation & Heterogeneity

Effects vary significantly by:

- Political identity strength
- Prior attitudes
- Demographics
- Context

Always segment results by:

- High vs low partisanship
- Age / education
- Baseline polarisation

Data Analysis Framework

Core metrics

- Mean change (pre–post)
- Effect size (Cohen's d)
- Difference-in-differences (treatment vs control)

Mechanism testing

- Correlate: change in mechanism → change in outcome

Durability analysis

- Plot outcome over time
- Calculate: % of effect retained

Experimental Learning Design

To improve interventions, build in A/B testing:

- Norms (Ingroup vs outgroup framing)
- Narratives (Personal story vs statistical)
- Contact (Structured vs unstructured)

Common Pitfalls (from the literature)

1. Only measuring attitudes → Misses behaviour
2. No follow-up → Misses decay
3. No mechanism measures → Cannot improve intervention
4. Weak control groups → Inflated results
5. Over-reliance on self-report → Limited real-world insight

Say No To Disinfo

Practical Evaluation Template for a pilot

Design

- Randomised (if possible)

Measures

- Affective polarisation (primary)
- 2 mechanism measures
- 1 behavioural proxy

Timeline

- Baseline
- Immediate
- 2 weeks

Sample size

- Large enough to detect small effects

Key Takeaways

Counter-polarisation interventions rarely fail completely, but they often produce small effects that fade quickly. So the role of evaluation is to:

1. Detect small changes reliably
2. Identify mechanisms
3. Optimise interventions
4. Test durability

A strong evaluation framework must:

- Combine outcomes + mechanisms + behaviour + implementation
- Track change over time
- Use comparison groups
- Be designed for learning, not just accountability