



Systematic Review

Interventions to Reduce Implicit Bias in High-Stakes Professional Judgements: A Systematic Review

Isabela Merla *, Fiona Gabbert and Adrian J. Scott

School of Mind, Body and Society, Goldsmiths, University of London, London SE14 6NW, UK;
f.gabbert@gold.ac.uk (F.G.); a.scott@gold.ac.uk (A.J.S.)

* Correspondence: m.merla@gold.ac.uk

Abstract

A systematic review was conducted to examine interventions designed to reduce the influence of implicit bias on professional judgements, with the aim of identifying strategies relevant to forensic and legal contexts. These decisions are often made under time pressure, ambiguity, and limited information, increasing reliance on intuitive judgement and mental shortcuts that can allow bias to shape how information is evaluated. Eight databases were searched and screened using predefined inclusion criteria. Studies were included if they assessed the behavioural impact of a bias-reduction intervention on decisions made by professionals or mock professionals in forensic, legal, healthcare, educational, or organisational settings. Thirty-eight studies met the inclusion criteria and were analysed. Interventions were mapped by mechanism, delivery format, and decision context. Systemic strategies, such as decision protocols, standardised rubrics, or changes to how information was presented, consistently outperformed individual-level approaches focused on changing attitudes or awareness. Effective interventions typically constrained discretion or embedded structured prompts at the point of judgement. However, most were tested in simulated settings, with limited evidence of long-term or applied effects. The review identifies strategies with the strongest empirical support and highlights those most effective, practical, and transferable to forensic and legal contexts.



Academic Editors: James Munro and Lee Curley

Received: 9 September 2025

Revised: 13 November 2025

Accepted: 14 November 2025

Published: 20 November 2025

Citation: Merla, I., Gabbert, F., & Scott, A. J. (2025). Interventions to Reduce Implicit Bias in High-Stakes Professional Judgements: A Systematic Review. *Behavioral Sciences*, 15(11), 1592. <https://doi.org/10.3390/bs15111592>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: implicit bias; decision-making; bias reduction; legal judgement; professional judgement; evidence-based practice

1. Introduction

1.1. Implicit Bias Within the Criminal Justice System

Implicit bias refers to automatic associations and stereotypes based on social characteristics such as race, gender and age, which can influence judgements and behaviour without conscious intent. It is a form of cognitive bias that arises specifically in social contexts, where cues related to social characteristics can affect how people evaluate or respond to others, even when they are motivated to make impartial decisions (Greenwald & Banaji, 1995). While the concept has been subject to debate and definitional variation (De Houwer, 2019; Gawronski et al., 2022; Holroyd & Sweetman, 2016), there is broad agreement that implicit bias can shape judgement and behaviour in ways that lead to discriminatory outcomes (Axt et al., 2025; Greenwald et al., 2022). This influence has been demonstrated across applied professional settings, including healthcare, education, employment, and particularly the forensic and legal context. This systematic review focuses on this latter topic, where the impact of bias on high-stakes decisions raises serious concerns about

fairness and equity. Decision-makers in the forensic and legal context may be especially susceptible to the effects of implicit bias because decisions are frequently made under conditions of time pressure, ambiguity, and limited information. These conditions often allow for considerable discretion and foster reliance on intuitive judgement and mental shortcuts, which increase the likelihood that automatic associations based on social characteristics will shape how information is processed and evaluated (Curley et al., 2022; Curley & Neuhaus, 2024; Dror, 2025).

Implicit bias has been shown to influence decision-making across multiple stages of the forensic and legal process, and among a wide range of professionals. For example, in the context of race, judges have been found to deliver harsher sentences to Black defendants than to White defendants, even when controlling for relevant legal variables (Mustard, 2001). More recently, a Ministry of Justice analysis of 21,000 indictable Crown Court cases found that, compared with White defendants, Black, Asian, and Chinese had approximately 53%, 55%, and 81% higher odds of imprisonment overall, and within drug offences ethnic minority defendants had around 240% higher odds of receiving immediate custody (Hopkins et al., 2016). Prosecutors are also more likely to charge Black individuals with offences that carry mandatory minimum sentences, which can lead to longer periods of incarceration and fewer opportunities for plea negotiations (Rehavi & Starr, 2014). Pretrial decisions reflect similar disparities, with Black defendants more likely to be assigned higher bail amounts or denied bail, which increases the likelihood of guilty pleas and results in more severe sentencing (Schlesinger, 2005; Dobbie et al., 2018).

Similarly, jurors may hold implicit beliefs about a defendant's credibility, guilt, or dangerousness based on race or other social characteristics, which can subtly shape how evidence is interpreted. These biases often reinforce prior assumptions rather than promote objective assessment. In particular, jurors who implicitly associate Black defendants with criminality are more likely to interpret ambiguous evidence as proof of guilt and give disproportionate attention to the defendant based on race, even in the absence of overt racial content (Pfeifer & Ogloff, 1991; Sommers & Ellsworth, 2001; Sargent & Bradfield, 2004; Mitchell et al., 2005; Young et al., 2014). Consistent with this, meta-analytic evidence shows small but reliable race-linked differences in mock-juror outcomes that typically reflect own-group favouritism. On average, White jurors are more punitive towards Black (and in some analyses, Hispanic) defendants, and Black jurors are more punitive towards White defendants (Mitchell et al., 2005).

Even outside the courtroom, implicit bias can affect how defence attorneys and death penalty lawyers advise clients, sometimes unconsciously adjusting their approach based on the client's race (Edkins, 2011). Forensic experts may also be influenced by racial cues, particularly in subjective or ambiguous cases, resulting in biased interpretation, presentation, or prioritisation of evidence (Dror et al., 2021).

1.2. Limitations of Current Intervention Approaches

Despite the serious consequences of implicit bias in forensic and legal contexts, very few interventions designed to mitigate its influence have been empirically tested, and most remain theoretically informed (Sah et al., 2015, 2016; Kovera, 2019). This is not only a gap in the criminal justice literature but also reflects a broader challenge across applied professional settings. In healthcare, for example, implicit bias contributes to disparities in diagnosis and treatment (FitzGerald & Hurst, 2017). Like forensic and legal contexts, it involves complex decision-making under time pressure and uncertainty; conditions that foster reliance on mental shortcuts and automatic associations. While numerous interventions have been trialled, most have shown little reliable impact on clinical decisions or behaviour (Vela et al., 2022).

A major reason for this lack of reliable impact lies in how existing implicit bias interventions have been designed and evaluated. Most established approaches focus on changing people's internal attitudes, associations, or awareness, and are typically evaluated based on whether they produce changes on psychological tests, most notably the Implicit Association Test (IAT), which measures the strength of automatic associations between social categories and evaluative traits based on reaction times (FitzGerald et al., 2019; Paluck et al., 2021). In much of the literature, reductions in IAT scores are taken as evidence that an intervention has been successful. Yet IAT scores are poor predictors of real-world behaviour, and improvements rarely translate into better decision-making or fairer outcomes (Forscher et al., 2019; Lai et al., 2016; Kurdi et al., 2019; Oswald et al., 2013; Buttrick et al., 2020). Moreover, most studies are conducted in laboratory settings, offering little insight into how interventions function in professional contexts (Greenwald et al., 2022).

This raises concerns about the practical utility of research to date. Applied settings, including the forensic and legal context, require strategies that reduce biased outcomes by improving decision quality and consistency, and by limiting the influence of implicit bias on professional judgement. Yet, most existing research provides limited guidance on how to achieve these outcomes in practice. A recent meta-analysis of methods for reducing prejudice highlights this gap, concluding that despite the wide range of strategies tested, there is still little reliable evidence identifying which interventions lead to meaningful behavioural change (Paluck et al., 2021). Without stronger evidence, it remains unclear which interventions should be prioritised for implementation in policy or adopted in professional contexts.

1.3. The Present Review

To address this gap, the present review synthesises empirical studies that have tested interventions designed to reduce the influence of implicit bias on consequential professional judgements. It focuses on decision-making contexts in which social characteristics such as race, gender and age are visible or can be inferred, and may shape evaluations in ways that create disparities. The review is designed to inform practice in forensic and legal contexts, and draws on research from fields where similar decision structures, challenges, and stakes are present to identify effective strategies in high-stakes environments and assess their potential for adaptation to legal and justice contexts.

To this end, the review will map the intervention landscape by identifying and describing the full range of strategies, specifying what each strategy changes, how it is delivered, and at what point in the decision process it operates. It will then evaluate the effectiveness of strategies by synthesising changes in decision outputs, focusing on the direction of effects, any impact on decision quality, and the durability of these effects where follow-up data are available. Finally, it will appraise the practicality and transferability of strategies by assessing whether interventions are feasible to implement within their original domain and whether their underlying mechanisms are likely to transfer to forensic and legal contexts, given their resource requirements, delivery conditions, time and operational constraints.

Three questions will be addressed. First, which intervention strategies have been tested to mitigate the influence of implicit bias in settings where evaluators make consequential decisions about other people? Second, how effective are these strategies when assessed against behavioural or evaluative outcomes, such as verdicts, treatment decisions, grades, shortlists, or hires? Third, which of these strategies can be adopted, scaled, and transferred to forensic and legal contexts, and under what conditions?

This review adds value in three ways. First, it focuses on behavioural and evaluative outcomes (i.e., what professionals do, decide, or recommend) rather than on changes in attitudes, such as shifts in belief, awareness, or self-reported bias, or changes in scores

on bias-related measures. While attitude change has often been used as a proxy for progress, there is now substantial evidence that changes in implicit or explicit attitudes may not reliably translate into changes in behaviour, and that behaviour can change in meaningful ways even when attitudes remain unchanged (Axt et al., 2025). By assessing interventions in terms of whether they directly reduce disparities in decisions, actions, and practices, this review reflects a broader shift in the field toward evaluating real-world impact. It contributes to a more applied and outcome-focused body of evidence, and supports the development of strategies that respond more directly to the practical demands of professional decision-making.

Second, this review links effectiveness to specific contexts in which interventions are implemented, recognising that they need to work outside of experimental or highly controlled environments. Real-world conditions often involve time pressure, limited or incomplete information, and other potential organisational constraints. Thus, by situating effectiveness within context, this review helps identify interventions that are feasible in practice, as well as offering important insights into the resources necessary to support their implementation.

Third, the review distinguishes between strategies that show potential for adaptation and transfer across professional domains and those whose effectiveness appears confined to specific settings, thus highlighting the broader question of transferability. In this conduct, the review supports decision-makers and practitioners in selecting effective strategies and understanding their scope and limitations, particularly within forensic and legal contexts. To our knowledge, it is the first review to systematically draw on evidence from multiple professional domains to examine and evaluate its applicability to forensic and legal contexts.

2. Method

A systematic review was conducted to identify effective intervention strategies that have been empirically tested to reduce the influence of implicit bias in professional settings where evaluators make consequential decisions about other people. This review was pre-registered on PROSPERO (CRD420251107033) and followed a pre-specified protocol developed in line with guidance for systematic evidence synthesis (Page et al., 2021).

The steps for constructing the systematic review involved a comprehensive search strategy, a structured screening process, data extraction, and a quality assessment of all included studies. Each of these stages is described below.

2.1. Search Strategy

In March 2025, a systematic search was conducted across eight electronic databases to identify eligible, published empirical articles. The databases were: APA PsycInfo, APA PsycArticles, Criminology Collection, ERIC, Social Science Database, ASSIA (Applied Social Sciences Index and Abstracts), PubMed, and Web of Science. These sources were selected to reflect the interdisciplinary scope of the review, covering psychology, social science, criminal justice, health, and education.

The search strategy combined terms across three conceptual domains: implicit bias, intervention, and decision-making. For bias, search terms included: implicit, unconscious, subconscious, automatic, heuristic, myth*, bias*, prejudic*, stereotyp*, attitude*, association*, discriminat*, and preference*. For interventions, terms included: debias*, intervention*, reduc*, training, strategy, method, program*, approach*, chang*, tool*, educat*, modif*, diminish*, counteract*, mitigat*, reframe*, and effective*. Decision-making terms included: decision-making, decision task, judg*, evaluat*, deliberation*, and verdict*. These terms were combined into search strings and adapted to the search architecture of

each database using Boolean operators and truncation symbols to maximise sensitivity. The full PsycInfo search string is provided in Appendix A.

Searches of titles, abstracts and keywords were conducted and were filtered to include only peer-reviewed journal articles published in English from 1 January 2000 onwards. This time frame was selected to reflect a documented shift in how bias and discrimination were conceptualised, including increased use of the terms ‘implicit bias’ and ‘unconscious bias’ in both scientific and public discourse (Greenwald et al., 2022). Studies from all geographical regions were considered, provided they met the inclusion criteria. The search was limited to studies involving adult human participants, and where possible, preprints, dissertations, and non-scholarly outputs were excluded to ensure that the knowledge obtained was peer reviewed. Full database-specific search strings, fields and filters are presented as a Supplementary Material (S1), reproducing the exact configurations used across all eight databases. Beyond the database searches, to ensure comprehensive coverage, additional studies were identified through backward and forward citation tracking, and review of reference lists from previous systematic reviews on related topics.

2.2. Screening Process: Inclusion and Exclusion Criteria

Studies were included if they met four primary criteria:

- (1) Focus on implicit bias: The study targeted bias in social evaluations related to social characteristics (e.g., race, gender, age). Bias was considered ‘implicit’ if it was described as automatic, unintentional, or unconscious, including related constructs using terms such as ‘automatic stereotyping’ or ‘unconscious associations’. Implicit bias was operationalised as a socially focused subtype of cognitive bias (i.e., a tendency for evaluative judgements to be influenced by automatic, unintentional processing of social characteristics). Accordingly, studies centred on non-social cognitive heuristics (e.g., anchoring, confirmation) or on explicit cultural attitudes/beliefs not characterised as automatic processes were ineligible and were excluded. Eligibility was therefore evidenced either by the authors’ framing of bias as automatic/unconscious social processing or by an empirical design that operationalised bias as a change in decisions produced by a social characteristic in the absence of task relevance or explicit intent (e.g., an evaluative decision task that manipulated a social characteristic while non-task-relevant to assess its effect on decisions).
- (2) Tested an intervention: The study examined an intentional strategy aimed at reducing or mitigating implicit bias in decision-making. Studies that reported incidental bias reduction without presenting it as a deliberate intervention were excluded.
- (3) Decision-making outcome: Outcomes had to reflect a meaningful judgement with real or simulated consequences about another person, such as sentencing, hiring, grading, performance evaluation, or treatment recommendation. Studies focused on attitudes, preferences, or affective ratings without an evaluative consequence were excluded.
- (4) Professional or mock-professional context: This included both real professionals acting in their formal roles (e.g., doctors, teachers, police officers) and lay participants who were explicitly instructed to take on a professional role (e.g., mock jurors, hiring decision-makers). Studies where outcomes could not be linked to an individual decision-maker were excluded.

In addition to the above, studies were excluded if they were non-empirical (e.g., literature reviews or theoretical articles), not published in English, and unpublished, including dissertations and preprints, to ensure peer-reviewed quality.

2.3. Screening Process: Title, Abstract, and Full Text Review

All articles retrieved through the database searches were exported into Excel, where duplicates were identified and removed. Titles and abstracts of the remaining articles were screened for relevance against the inclusion criteria. Full texts were then retrieved for all articles that appeared to meet the inclusion criteria or for which eligibility was uncertain based on the abstract.

Screening was conducted in two phases. In the first phase, the first author screened the titles and abstracts of all identified articles. If the first author was unsure about the inclusion of any studies, this was discussed further with the Research Team. In the second stage, the full texts of potentially eligible articles were retrieved and assessed against the full inclusion criteria by the first author and checked by the second author. Discrepancies were resolved through discussion and reference to the protocol, with input from the third author where necessary. Reasons for exclusion at the full-text stage were recorded.

The screening process adhered to PRISMA guidelines. Figure 1 presents a PRISMA flow diagram that summarises the number of records identified, screened, assessed for eligibility, and included in the final review. In total, 38 studies from 26 articles met the inclusion criteria and were included in this review.

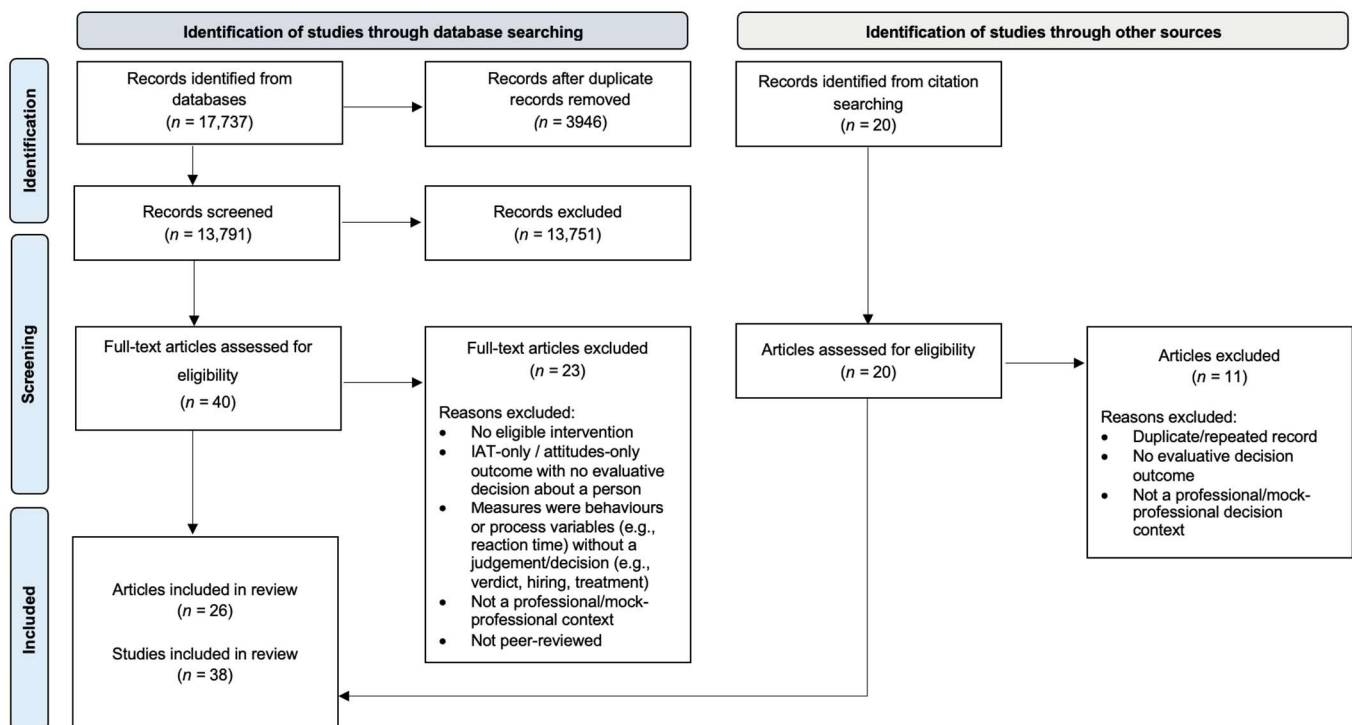


Figure 1. PRISMA flow diagram summarising the literature searching and sifting process.

2.4. Data Extraction

A Searchable Systematic Map (SSM) was developed in Excel to represent the 38 studies that met the review's inclusion criteria. Each study was systematically coded using a structured framework designed to support cross-study comparison, analysis, and inform the narrative synthesis. The SSM functioned as both a descriptive summary and an analytical tool, allowing the evidence to be filtered, examined, and compared across key dimensions to explore what worked, how it worked, and how likely it was to be transferable to applied professional settings.

Data extraction was guided by four domains: 1. Study details, 2. Intervention characteristics, 3. Outcomes and findings, and 4. Delivery practicality and implementation feasibility. Full definitions and coding rules are provided in Appendix B.

- (1) **Study details:** Captured core methodological and contextual information (e.g., year, country, professional field, study design, participant role, paradigm). Three ecological indicators (sample realism, task realism, and context realism) were also coded to assess the ecological validity of each study and the extent to which findings might generalise to applied professional settings.
- (2) **Intervention characteristics:** Documented each intervention's design, delivery, mechanism of action, and timing. Interventions were classified both by level of operation (individual, systemic, or mixed) and by mechanism (e.g., altering information available at judgement, adding structure to reduce discretion, prompting self-regulation, reframing assumptions, or targeting automatic associations). These classifications facilitated structured comparison of strategies and their potential transferability.
- (3) **Outcomes and findings:** Extracted evidence on whether interventions reduced bias in consequential decisions (e.g., sentencing, hiring, grading, treatment). Effectiveness was judged against baseline bias and assessed for statistical significance, consistency, and durability. Secondary outcomes (e.g., implicit bias measures, participant feedback) were also recorded to provide contextual insight into mechanisms and broader impact.
- (4) **Delivery practicality and implementation feasibility:** Recorded information on format, materials, time demands, training needs, and scalability. Interventions were appraised for practicality (resource and process requirements) and feasibility (likelihood of adoption, fidelity, and sustainability under real-world constraints). Ratings of high, moderate, or low were applied across key aspects such as cost, facilitation needs, duration, and scalability.

Together, these domains allowed the review to assess not only whether interventions were effective, but also how they operated and how feasible they would be to implement in professional practice, particularly in high-stakes forensic, legal and related contexts. For more information about the data extraction process, please see Appendix B.

2.5. Quality Assessment

A structured critical appraisal was conducted to assess the internal validity, methodological quality, and reporting transparency of all included studies. The purpose was to strengthen the reliability of the synthesis by identifying the strength of evidence behind each reported effect and to avoid over-weighting findings from studies with design limitations. Appraisal was not used to exclude studies unless fundamental flaws affected meaningful interpretation.

Appraisal followed the Quality Assessment Tool for Diverse Studies (QuADS), a published tool developed for systematic reviews, selected for its applicability across diverse empirical designs (Harrison et al., 2021). The tool covers 13 domains, including clarity of theoretical framework, appropriateness of study design and sampling strategy, and transparency in data collection, recruitment, analysis methods and stakeholder involvement. Each domain was scored from 0 (not at all) to 3 (fully), based on information reported in the publication. Following QuADS guidance, domain scores were interpreted qualitatively and synthesised into an overall judgement of low concern (most domains scored 2–3), some concerns (mixed scoring patterns with one or more domains scored at 1), or high concern (multiple domains scored 0–1), regarding methodological and reporting adequacy, based on predefined criteria. Full domain descriptions and scoring anchors are provided in Appendix C.

3. Results

3.1. Overview of Study Characteristics

The review includes 26 articles reporting 38 distinct studies. Studies were published between 2002 and 2024, with the majority (73.7%) published from 2020 onwards and half (50.0%) between 2021 and 2024. Earlier studies were infrequent, with a smaller number of articles published between 2002 and 2019 (26.3%), reflecting a recent increase in applied debiasing research. Most studies were conducted in the United States (63.2%), followed by Western Europe (26.3%) and Asia (10.5%).

Most studies investigated bias in workplace or organisational settings (63.2%), followed by criminal justice (13.2%), healthcare (13.2%), education (7.9%), and civic or political decision-making (2.6%). The most targeted forms of implicit bias were race/ethnicity (42.1%) and gender/sex (57.9%). Age was included in fewer studies (10.5%), followed by socioeconomic status (7.9%). A small subset of studies (15.8%) investigated more than one social characteristic within the same task or across separate comparisons. Interventions that were investigated were designed to operate at the individual-level (47.4%) and the systemic-level (47.4%) in equal measure, with a small number combining both (5.3%). Online delivery was more common (63.2%) than in-person delivery (36.8%).

Ecological realism was mixed. Sample realism was high in around two-fifths of studies (39.5%), where participants were practising professionals or professionally relevant trainees. It was low in a considerable proportion (36.8%), typically involving general student populations or online panels with limited applied relevance, and moderate in a smaller number (23.7%), which involved working adults or students with some professional relevance. Task realism was predominantly moderate (55.3%), characterised by simplified but recognisable approximations of real decision scenarios. It was low in some cases (34.2%), with abstract or overly artificial formats, and high in a few studies (10.5%) where tasks closely reflected the complexity and structure of high-stakes decision-making. When considering sample, task, and paradigm together, overall context realism was low in half of the studies (50.0%), followed by moderate (39.5%), and high in only a small subset (10.5%). In practical terms, most of the interventions were tested on simplified versions of professional decisions within lower-stakes, survey-based environments.

The timing of intervention delivery was typically close to the decision point. Interventions were delivered immediately before the decision in half of the studies (50.0%) and during the decision process in a smaller subset (44.7%). Only a small number (5.3%) introduced the intervention earlier, at a greater temporal distance from the judgement. Durability was rarely assessed. Only a minority of studies (15.8%) included any follow-up measurement, with time intervals ranging from one week to approximately ten months. The remainder (84.2%) measured outcomes only immediately after the intervention. Studies providing conclusions about the persistence of intervention effects were therefore limited.

3.2. Overview of Intervention Effectiveness

The findings show a clear pattern of intervention effectiveness by intervention approach. Systemic-level strategies, which target the decision environment, accounted for most strong effects—whereby effects were labelled as strong when an intervention produced a clear, sample-wide main-effect reduction on the targeted decision outcome, moderate when improvements were conditional (e.g., limited to subgroups, specific measures or interaction-only patterns) rather than uniform across the full sample, and limited when there was no effect on the targeted outcome or when the study lacked a baseline disparity on that outcome. Of the eighteen systemic-level studies identified, fourteen produced strong effects (77.8%). Contrastingly, individual-level strategies, which target the decision-maker, showed strong effects in seven of eighteen studies (38.9%), while mixed-level interventions,

combining changes to individual decision-making with structural adjustments, produced mostly conditional or limited evidence. These patterns are examined in detail below.

3.3. Systemic-Level Interventions

Systemic strategies produced the most consistent improvements, accounting for two-thirds of all strong findings in this review (66.7%). These interventions were organised into two recurrent mechanisms: 1. Altering information available at judgement; and 2. Adding structure that limits discretion.

3.3.1. Altering Information Available at Judgement

Ten studies examined interventions that changed what decision-makers saw at the decision point or how this information was presented (26.3% of all studies; 55.6% of systemic studies). Seven studies within workplace contexts examined gender (71.4%) or ethnicity (28.6%) in shortlisting or hiring tasks (Feng et al., 2020). Two studies targeted sex/gender and race/ethnicity in professional screening (Friedmann & Efrat-Treister, 2023; Pershing et al., 2021), and one examined gender bias in political committee selection (Wall et al., 2022). Overall, seven of ten interventions were effective (70.0%).

Changing what candidates were shown at the decision point was the most consistently effective strategy. All seven shortlisting and hiring studies produced more balanced shortlists and, when a single choice was required, increased the likelihood that a woman or minority candidate was selected (Feng et al., 2020). Partitioning candidates by social category encouraged selectors to distribute their selections across categories and outperformed brief prompts that only stated category information, indicating that the display design, rather than information alone, drove the effect. Grouping also altered selections. In these tasks, profiles from one category were shown together while profiles from the comparison category remained listed individually. This layout shifted attention to the individually listed profiles. Grouping the majority increased the selection of minority candidates, whereas grouping the minority reduced it. Where assessed, these gains did not reduce decision quality, though stronger implicit biases (higher gender IAT scores) weakened effects. By contrast, interventions that modified identity cues to prevent stereotype activation or added process transparency to encourage more proportionate exploration produced limited results. In STEM hiring, standardising an availability cue by adding the same 'long-hours' note to matched male and female CVs, in the form of a generic statement of willingness to work extended hours, shifted how managers weighed that criterion but not hiring likelihood (Friedmann & Efrat-Treister, 2023). In residency screening and political committee formation, there was no measurable change from controls (Pershing et al., 2021; Wall et al., 2022).

These studies offered limited evidence for real-world impact. Most evaluations used non-professional samples and simplified tasks. Only one shortlisting study used an in-person paper-resume procedure and reproduced the display effect (Feng et al., 2020). Residency redaction and STEM screening involved practising faculty or managers, yet outcomes did not differ between the compared groups, leaving no baseline disparity against which to assess change. None of the studies included follow-up measurement; however, for systemic strategies, the central issue is not retention, but whether protocols are implemented with high fidelity and without unintended consequences such as accuracy errors. Future replications with professional decision-makers should include fidelity checks and assess decision quality and error rates, alongside equity outcomes, to establish suitability for high-stakes practice.

3.3.2. Adding Structure That Limits Discretion

Eight studies tested interventions that constrained bias-prone discretion by adding structure to decision processes (21.1% of all studies; 44.4% of systemic studies). Six studies focused on workplace hiring or shortlisting (75.0%) (Bragger et al., 2002; Uhlmann & Cohen, 2005; Lucas et al., 2021), one on education grading (12.5%) (Quinn, 2020), and one on clinical care management (12.5%) (Hamm et al., 2020). Most interventions targeted gender (75.0%), with race/ethnicity considered in two studies (25.0%). Seven out of eight interventions were effective (87.5%).

Interventions that required decision-makers to follow standardised procedures or predefined evaluative criteria produced the most consistent improvements in outcomes. In healthcare, a structured labour-management protocol reduced caesarean rates and improved neonatal health for Black patients, without adverse effects for others (Hamm et al., 2020). In education, grading with a predefined, detailed rubric eliminated racial disparities that were otherwise observed under more subjective conditions (Quinn, 2020).

Hiring studies applied comparable strategies to constrain discretion in candidate evaluations. Here, structured interviews with behaviourally anchored rating scales reduced subjectivity and minimised stereotype-driven assessments (Bragger et al., 2002). Another intervention asked evaluators to assign weights to selection criteria before reviewing applications. This pre-commitment eliminated the gender preference typically shown by male raters (Uhlmann & Cohen, 2005). Three other hiring studies increased shortlist length by requiring evaluators to identify six candidates rather than three, which increased the number of women but found no corresponding change in final selections (Lucas et al., 2021).

These studies offer strong evidence that standardised protocols and predefined criteria can improve decision-making by limiting the role of discretionary judgement. Most of the hiring studies using this mechanism showed similar promise but often relied on student or online samples using simplified tasks, leaving generalisability to high-stakes hiring decisions unclear. Similar to the previous mechanism of systemic interventions, none of the studies assessed outcomes beyond the immediate decision or evaluated decision quality. Conclusions about fidelity use and error rates are, therefore, limited and should be investigated by future research.

3.4. Individual-Level Interventions

Individual-level strategies contributed to one-third of all strong findings in this review (33.3%). These interventions were organised into three categories: 1. Prompting self-regulation at the point of decision, 2. Reframing assumptions, and 3. Targeting automatic associations.

3.4.1. Prompting Self-Regulation at the Point of Decision

Nine studies encouraged individuals to pause, reflect, or engage in corrective routines before making a judgement (23.7% of all studies; 50.0% of all individual studies). Most were conducted in workplace evaluation or selection contexts (44.4%) (Anderson et al., 2015; Döbrich et al., 2014; Kleissner & Jahn, 2021), followed by criminal justice decision-making (44.4%) (Lynch et al., 2022; Ruva et al., 2024; James et al., 2023), and school disciplinary decisions (11.1%) (Naser et al., 2021). Targeted biases included race/ethnicity (55.6%), gender (44.4%), age (33.3%), and socioeconomic status (11.1%). Three of these studies were effective (33.3%).

The clearest evidence for real-world impact came from a field-based policing study that combined classroom instruction with high-fidelity simulation training. Officers learned about how bias can influence judgement, concrete practices for fairness and de-escalation, and then practised in simulators with immediate feedback. This study produced sustained

improved performance in real-world interactions, particularly with community members low in socioeconomic status, and reduced discrimination-related complaints over the 10-month period of monitoring (James et al., 2023).

Consistent improvements also came from brief prompts embedded directly into the decision workflow, designed to encourage reflection and direct attention toward relevant evaluation criteria. In workplace hiring, an on-screen reminder that age can bias evaluation shifted attention to job-relevant skills, which narrowed or eliminated age gaps in ratings without reducing attention to applicant qualifications (Kleissner & Jahn, 2021). A similar intervention in education added a ‘pause and plan’ step to school discipline procedures. This lowered teachers’ referral rates for Black students and shifted perceptions of those students’ behaviour, suggesting a move toward more deliberate judgement (Naser et al., 2021). Some improvements were also observed in two HR studies, where short age-bias warnings eliminated penalties against older applicants in both performance appraisal and hiring tasks. However, accountability requirements helped older men more than older women, pointing to selective effects across intersecting identities (Döbrich et al., 2014).

By contrast, prompts lacking clear behavioural guidance or when baseline disparities were absent showed weaker results. Two jury studies tested implicit-bias instructions and orientation videos, finding these increased bias-related discussion during deliberation, but not verdict disparities, and sometimes backfired and encouraged overcorrection (Lynch et al., 2022; Ruva et al., 2024). Similarly, in workplace leader evaluations, prompts produced inconsistent results that were moderated by raters’ implicit attitudes (Anderson et al., 2015).

Ecological realism varied widely, with most studies relying on simplified decision tasks, limiting what can be inferred about generalisability to high-stakes settings. An exception was that of James et al. (2023), which embedded training into professional routines and demonstrated that self-regulation works best when paired with clear behavioural strategies and feedback. Where individual differences were tested, effects were uneven, raising questions about scalability and equity across intersectional identities.

3.4.2. Reframing Assumptions

Five studies tested interventions that reshaped how decision-makers construe people and attribute capabilities (13.2% of all studies; 27.8% of individual studies). Three were conducted in workplace evaluation or selection contexts (60.0%) (Derous et al., 2021; Liu et al., 2023), followed by two in healthcare treatment decisions (40.0%) (Hirsh et al., 2019; Neal et al., 2024). Targeted biases included race/ethnicity (40.0%), gender (40.0%), age (20.0%) and socioeconomic status (20.0%). Three interventions were effective (60.0%).

Across studies, the most consistent effects came from interventions that explicitly surfaced bias in decision-makers’ own judgements and redirected them toward individuated, evidence-based appraisals rather than group-based assumptions. In healthcare, one intervention delivered real-time feedback on treatment decisions with virtual perspective-taking modules, increasing empathy and equity in pain management (Hirsh et al., 2019). A separate intervention reframed bias as a shared responsibility and paired it with guidance on shared decision-making, reducing age-based assumptions in cancer care (Neal et al., 2024).

In a workplace evaluation, one study tested two training formats aimed at promoting more individualised judgement of Arab/Moroccan applicants. One showed realistic workplace misunderstandings between majority and minority employees, after which participants made judgements, received corrective feedback, and then engaged in a debrief session with discussion and role-play. This approach built more nuanced mental models of cultural difference and improved job suitability ratings. The second task asked participants to study scripted workplace vignettes, then recall specific behaviours into key performance domains before rating the target. This also improved ratings, though effects were smaller.

For both, improvements faded by three months, highlighting the need for reinforcement (Derous et al., 2021). By contrast, two studies used short reframing prompts to shift assumptions about who is seen as a ‘leader’. Both introduced a universal framing, emphasising that leadership potential is widespread and developable, and tested its effects on gender bias in candidate evaluations. Effects were mixed and, in one case, no bias was observed under control conditions, limiting conclusions (Liu et al., 2023).

Overall, healthcare interventions aligned most closely with real-world practice, while workplace studies often relied on online evaluations by students or adults with varied experience. Durability remained limited, with effects fading without reinforcement.

3.4.3. Targeting Automatic Associations

Four studies attempted to directly alter implicit associations (10.0% of all studies; 22.2% of individual studies), three in workplace hiring or evaluation (75.0%) (Kawakami et al., 2005, 2007; Brauer & Er-rafy, 2011), and one study in criminal justice (25.0%) (Salmanowitz, 2018). Bias targets were split between gender (50.0%) and race/ethnicity (50.0%). One intervention was effective (25.0%).

Within these studies, the consistent effect came from an intervention that reshaped automatic associations by increasing perceived variability within social groups, thereby reducing reliance on stereotypes at the point of decision. In a hiring simulation, participants prompted to complete simple variability sentences about Arab individuals (e.g., “Whereas some. . . , other. . .”) showed no bias in composite ratings, rankings, or interview selections, unlike participants exposed to homogeneity prompts or controls. Mediation analyses showed that the manipulation increased perceptions of within-group variability (i.e., diversity), and this increase accounted for the reduction in biased evaluations (Brauer & Er-rafy, 2011).

By contrast, counter-stereotype retraining did not reliably reduce gender bias in candidate selection, and sometimes overcorrected (Kawakami et al., 2005, 2007). In criminal justice, a brief VR task that placed participants in a Black avatar reduced race-IAT scores and increased confidence in not-guilty verdicts. However, no baseline disparity was present in the control group, limiting what can be inferred about bias reduction (Salmanowitz, 2018).

Ecological realism was generally low: all four studies used laboratory-style simulations with student or community samples. Follow-ups were rare and inconclusive. Overall, targeting automatic associations shows proof of concept for altering representations, but evidence of durable impact in real-world settings is lacking.

3.5. Mixed-Level Interventions

Mixed strategies were rare, appearing in only two studies (5.6%). One addressed race in clinical interactions (Dahlen et al., 2024), the other socioeconomic status and ethnicity in educational evaluations (Lehmann-Grube et al., 2024).

Neither produced strong effects, but suggested promising avenues for integrating skills-based regulation with low-cost process prompts. The healthcare study combined brief pre-learning with high-fidelity simulation to train implicit bias mitigation strategies such as individuation, partnership building and perspective-taking, reinforced through immediate feedback and reflective debriefs. Modest improvements in judgement scores and some unit-level indicators, such as a decrease in security dispatch calls during and after the intervention, were observed immediately and at three-month follow-up. However, results were descriptive and lacked a control group. In the education study, a short online module combined theory-based instruction with two process constraints: slowing down and applying criteria, and an ‘if-then’ intention before each decision. Teacher judgements

of low socioeconomic students improved, particularly for academic capability. However, other effects were small or inconsistent.

3.6. Practicality of Interventions

Across the 38 studies reviewed, only two interventions (5.3%) demonstrated strong effects on decision outcomes in applied professional settings. These differed both in intervention approach, with one systemic and one individual, and in their practicality, reflecting broader patterns whereby systemic interventions were generally easier to implement.

The first intervention involved a standardised protocol that reduced racial disparities amongst practising clinicians (Hamm et al., 2020). The second was a police training intervention that improved real-world behaviour and reduced discrimination-related complaints over approximately 10 months (James et al., 2023). The clinical protocol required only modest investment for development and monitoring and no additional time at the point of care, making it relatively low in cost and high in duration practicality (Hamm et al., 2020). In contrast, the policing programme required about 12 h of delivery time, access to simulation facilities, and trained instructors, resulting in low practicality for cost, facilitation and duration (James et al., 2023). Together, these two studies offer rare examples of strong impact in applied professional settings, suggesting that effective interventions exist across different delivery demands: protocols suit contexts with limited training time or facilitation capacity, whereas classroom learning and simulation-training programmes are more suitable for organisations able to invest in extended delivery time, trained facilitators, and access to specialised equipment.

Among effective studies conducted under lower ecological realism, systemic-level strategies were the most practical to deliver and implement. Here, thirteen effective systemic interventions were highly practical (100.0%) relying on simple adjustments to decision flows, such as grouping candidates in their display (Feng et al., 2020), extending shortlists (Lucas et al., 2021), or predefining criteria and including structure (Bragger et al., 2002; Quinn, 2020). These could be delivered in seconds to minutes without facilitators, all fitted within existing workflows with negligible marginal cost, and none required specialist involvement for delivery.

Effective or promising individual-level approaches were also practical when they were brief and embedded directly in the workflow. Across seven effective individual strategies, four were also highly practical (57.1%). These included anti-bias or self-regulation prompts (Kleissner & Jahn, 2021; Naser et al., 2021) and simple accountability or warning prompts (Döbrich et al., 2014), which were mostly self-guided and required no facilitation. In contrast, the remaining three interventions (42.9%) were less practical. These included intercultural or structured recall training (Derous et al., 2021) and a clinician module that combined real-time treatment feedback with perspective-taking videos (Hirsh et al., 2019). These interventions were longer and required either trained facilitators or dedicated delivery platforms. Overall, practicality declined as dependence on trainers, specialised environments, or lengthy formats increased.

3.7. Transferability of Interventions

The most effective interventions varied not only in delivery demands but also in how easily their strategies could be transferred across domains and professional settings. While practicality describes how feasible an intervention is to deliver in its original format, transferability reflects whether that intervention could plausibly work in other contexts or be adapted to different organisational structures. While many effective interventions have features that support transfer, only a minority were tested outside their original setting, and

even fewer assessed whether effects held under different conditions or over time. Therefore, suggestions are made with caution.

Among the few studies showing strong real-world effects, only one clearly offered a transferable potential. The standardised clinical protocol introduced in obstetrics was integrated into routine care and operated through pre-specified thresholds and actions, making it adaptable to any context where decision points are standardised and outcomes are routinely monitored (Hamm et al., 2020). The underlying mechanism, specifically removing discretion through rule-based structures, applies broadly wherever decisions follow a predictable structure or are repeated. In contrast, the police training programme that combined instruction with simulator exercises was effective but less transferable, as it relied on dedicated equipment and trained facilitators, limiting use to settings with similar resources (James et al., 2023). The core idea of skills-based practice with feedback could generalise, but only with substantial investment or simplification.

Transferability was also strong among effective systemic strategies tested under lower ecological realism. These relied on structural features common across decision-making contexts, such as shortlists, application forms, or performance rubrics, and worked by changing how information was presented or how choices were constrained (Feng et al., 2020; Lucas et al., 2021; Uhlmann & Cohen, 2005). Because they are embedded in standard formats and require little facilitation, they could be easily utilised across settings and implemented through policy changes, interface updates, or template modifications. In principle, these mechanisms are deployable to all the contexts where evaluators compare or assess individuals under conditions of discretion and potential ambiguity.

In contrast, individual strategies showed more uneven potential for transfer. Prompts that are brief, self-guided, and delivered at the decision moment are broadly applicable across settings. For instance, a short screen reminder about age bias could be added to any hiring or appraisal interface with no structural modifications (Kleissner & Jahn, 2021). Similarly, reflective cues and pause instructions fit naturally into decision workflows in education and HR and could transfer with only minor contextual tailoring (Naser et al., 2021; Döbrich et al., 2014). However, interventions that rely on facilitators, video debriefs, or customised platforms are less transferable because they require specific technical infrastructure, such as interactive feedback systems or high-fidelity simulations (Hirsh et al., 2019). In these cases, the mechanism of the interventions could still be transferable, but the format of the delivery would require major redesign.

3.8. Critical Appraisal

All included studies were critically appraised using the QuADS tool. Consistent with guidance, scores were used to interpret confidence in findings and to highlight where evidence was strongest or required caution, rather than to exclude studies or impose numerical weights. Domain-level QuADS scores were recorded for each study and used to identify common strengths and weaknesses across the evidence base. The ratings were not intended to compare or rank study quality, but to clarify where methodological or reporting issues were most evident and how these shaped confidence in overall conclusions. This structured approach ensured consistency across studies of differing designs and helped identify recurring methodological and reporting patterns. Table 1 presents a QuADS summary of domain ratings (mean, standard deviation and range), accompanied by brief interpretative summaries.

Across the 38 studies, overall reporting quality was sufficient to support synthesis. Importantly, no study raised high concern, while ten indicated strong reporting and design clarity (Derous et al., 2021; Friedmann & Efrat-Treister, 2023; Hamm et al., 2020; Hirsh et al., 2019; James et al., 2023; Lehmann-Grube et al., 2024; Lynch et al., 2022; Quinn, 2020;

[Ruva et al., 2024](#)). The remaining 28 presented appropriate designs and analyses, but raised some concern in at least one domain, most often sampling justification, stakeholder involvement, or procedural transparency. Therefore, these were interpreted with greater caution, specifically with regard to their generalisability to applied professional settings.

Table 1. QuADS summary of domain ratings and interpretative summaries.

QuADS Domain	Mean	SD	Range	Interpretive Summary
1. Theoretical or conceptual underpinning to the research	2.97	0.16	2–3	Frameworks were explicitly stated and applied through design and outcomes.
2. Statement of research aim/s	2.97	0.16	2–3	Aims were clearly and explicitly stated in all studies.
3. Clear description of research setting and target population	2.58	0.50	2–3	Settings/populations were described, though contextual detail was often limited.
4. The study design is appropriate to address the stated research aim/s	2.32	0.47	2–3	Designs matched aims but sometimes relied on simplified formats.
5. Appropriate sampling to address the research aim/s	1.71	0.80	1–3	Most lacked strong sampling justification (power, representativeness, recruitment detail).
6. Rationale for choice of data collection tool/s	2.50	0.51	2–3	Most studies provided a rationale for chosen instruments but lacked psychometric validation details.
7. The format and content of the data collection tool are appropriate to address the stated research aim/s	2.58	0.50	2–3	Tools were appropriate and clear; many were author-designed without psychometrics.
8. Description of data collection procedure	2.84	0.37	2–3	Procedures were clearly described; minor omissions occurred.
9. Recruitment data provided	2.05	0.93	1–3	Recruitment/attrition reporting was uneven and often incomplete.
10. Justification for the analytic method selected	2.76	0.63	1–3	Analytic choices were usually well justified and linked to aims.
11. The method of analysis was appropriate to answer the research aim/s	2.84	0.44	1–3	Analyses were generally suitable for the study design and data type.
12. Evidence that the research stakeholders have been considered in the research design or conduct.	0.68	0.74	0–2	Stakeholder engagement was minimal, noted only in a small subset of applied or participatory studies; substantial involvement was absent.
13. Strengths and limitations critically discussed	2.45	0.55	1–3	Most studies provided reflective discussion of limitations, though the depth varied.

Sampling was the most recurrent weakness. Only eight studies (21.1%) provided detailed justification, aligning samples with the study aims and/or offering power arguments ([Friedmann & Efrat-Treister, 2023](#); [Hamm et al., 2020](#); [Lynch et al., 2022](#); [Quinn, 2020](#); [Ruva et al., 2024](#); [Döbrich et al., 2014](#)). Eleven studies (28.9%) offered partial justification, typically linking sample characteristics with aims ([Dahlen et al., 2024](#); [Derous et al., 2021](#); [Hirsh et al., 2019](#); [James et al., 2023](#); [Lehmann-Grube et al., 2024](#); [Liu et al., 2023](#); [Naser et al., 2021](#); [Pershing et al., 2021](#)), while the remaining 19 studies (48.5%) provided only minimal sampling information ([Anderson et al., 2015](#); [Bragger et al., 2002](#); [Brauer & Er-rafiy, 2011](#); [Feng et al.,](#)

2020; Kawakami et al., 2005; Kawakami et al., 2007; Kleissner & Jahn, 2021; Neal et al., 2024; Salmanowitz, 2018; Wall et al., 2022). Despite a focus on professional decision-making, many studies either did not prioritise or did not document sample realism, recruitment transparency, or power planning. In the narrative synthesis, these sampling limitations were used to weight interpretations toward studies with transparent, appropriate sampling and higher ecological realism.

Stakeholder engagement was also limited overall. No study showed considerable stakeholder involvement. Six studies (15.8%) showed moderate evidence, including piloting that informed design or involvement by domain experts (Dahlen et al., 2024; Friedmann & Efrat-Treister, 2023; Hamm et al., 2020; Hirsh et al., 2019; James et al., 2023; Pershing et al., 2021). Fourteen (36.8%) evidenced minimal involvement, for example, limited piloting (Bragger et al., 2002; Deros et al., 2021; Feng et al., 2020; Lynch et al., 2022; Naser et al., 2021; Neal et al., 2024; Ruva et al., 2024). The remaining 18 studies (47.4%) reported no stakeholder involvement (Anderson et al., 2015; Brauer & Er-rafiy, 2011; Kawakami et al., 2005, 2007; Kleissner & Jahn, 2021; Lehmann-Grube et al., 2024; Liu et al., 2023; Quinn, 2020; Salmanowitz, 2018; Wall et al., 2022; Döbrich et al., 2014). Therefore, evidence on feasibility, acceptability, and fit with existing professional workflows was sparse.

Measurement choices were typically sensible for isolating social-characteristic cues, but many tasks and instruments were author-designed rather than validated scales or established scenarios. This is not inherently problematic, as custom materials allowed several studies to hold task content constant across conditions (e.g., Anderson et al., 2015; Deros et al., 2021; Quinn, 2020). However, this introduced heterogeneity in how implicit bias was operationalised and detected, and psychometric properties (e.g., reliability, validity, invariance) were rarely reported. In addition, a small subset of seven studies (18.4%) reported absent or atypical baseline disparities in their primary decision outcomes, limiting any meaningful conclusion about bias reduction (Liu et al., 2023; Lynch et al., 2022; Pershing et al., 2021; Ruva et al., 2024; Salmanowitz, 2018; Wall et al., 2022). Although a lack of baseline bias is not, in itself, a threat to internal rigour, it does suggest that the stimuli or outcome measures may have lacked diagnostic sensitivity to elicit or detect bias.

4. Discussion

This review synthesised 38 studies that tested interventions intended to reduce the influence of implicit bias on consequential judgements. Across the evidence base, a consistent pattern emerged between intervention approaches, suggesting that changing the decision environment was more effective than trying to change decision-makers. Among systemic-level interventions, 14 were effective (77.8%), compared to only seven (38.9%) individual-level interventions. The two mixed studies yielded only conditional or limited evidence. Of the 21 studies that produced strong outcomes, two-thirds (66.7%) were achieved by systemic changes.

These results align with the growing critique around individual-level bias training. While earlier implicit bias interventions have often targeted implicit associations or awareness, there is now a broad consensus that these changes rarely translate into behavioural improvements in applied settings (Axt et al., 2025; Forscher et al., 2019; FitzGerald et al., 2019; Paluck et al., 2021). This review supports this view and points to a more practical conclusion: in high-discretion fields like forensic and legal contexts, where time pressure, ambiguity and variability are common constraints (Curley et al., 2020; Dror, 2025), interventions should prioritise the design of the decision environment, with individual strategies used selectively to support key skills or processes.

Most of the systemic strategies worked through two mechanisms. The first was altering what information was visible at the point of judgement. Here, simple display changes,

such as partitioning candidates or grouping options, shifted attention and changed which comparisons were made without impacting decision quality, while increasing diversity of choice for women and minority candidates (Feng et al., 2020). The second mechanism was adding structure that limits discretion through predefined criteria, standardised rubrics or protocols (Bragger et al., 2002; Hamm et al., 2020; Lucas et al., 2021; Quinn, 2020; Uhlmann & Cohen, 2005). These interventions prevented evaluators from unconsciously shifting standards based on identity cues and made decisions more consistent and less reliant on subjective interpretation. Taken together, the most effective strategies intervene directly at the point of decision to constrain the influence of social categories and direct judgement towards evidence-based criteria.

This has clear relevance for contexts where redacting identifying details is not always feasible or desirable, including forensic and legal contexts. Pre-committing to legally relevant criteria before seeing identifying information could reduce shifting standards in complex judicial decisions, whereas structured decision aids, such as behaviourally anchored rubrics, could support any decisions that involve multiple, observable case-linked elements (e.g., credibility assessments, sentencing recommendations). Therefore, systemic strategies could aid in legal or forensic judgements where multiple people or claims must be considered at once, through structuring the presentation of information to reduce salience asymmetries and encourage comparisons on case-relevant grounds rather than intuitive or identity-based cues. Importantly, these strategies do not require evaluators to change their beliefs or reasoning style, only to operate within a process that improves the consistency and traceability of judgement.

Where decision quality was evaluated, improvements did not appear to coincide with reductions in accuracy, though this was rarely assessed. Going forward, implementation should include monitoring for fidelity, error rates, and unintended effects. Systemic strategies rely on consistent use to be effective; however, they can be embedded in forms, protocols, and digital systems that allow close monitoring. This also makes them attractive from a governance perspective, offering tools not just for reducing bias, but for improving transparency and accountability.

The findings on individual-level interventions were more mixed. Where effective, these strategies tended to include both a corrective goal and an opportunity to practice decision skills under realistic or time-pressured conditions. This was best illustrated in a field study of police officers, where a training programme combining classroom instruction with simulation improved real-world behaviour and reduced complaints over a sustained period (James et al., 2023). Other promising examples included short prompts embedded into decisions, designed to direct attention to relevant criteria or encourage reflection, with effects observed on selection, appraisal, and disciplinary outcomes (Kleissner & Jahn, 2021; Naser et al., 2021; Döbrich et al., 2014). Additionally, three studies (Derous et al., 2021; Hirsh et al., 2019; Neal et al., 2024) used structured reframing interventions to prompt evidence-based evaluation, increase empathy, or reduce reliance on group-based cues, with observed effects on hiring, healthcare, and treatment decisions. Thus, the effective individual-level strategies worked not by changing underlying attitudes, but by interrupting automatic judgement, increasing motivation to individuate and supporting more effortful, evidence-linked reasoning.

By contrast, most interventions focused solely on awareness or association retraining showed little impact. Awareness-only prompts that lacked behavioural guidance sometimes triggered overcorrection, as in jury studies where participants misinterpreted justified scepticism as bias (Lynch et al., 2022). Some prompts produced uneven effects across intersecting identities, helping some subgroups but not others (Döbrich et al., 2014). In other cases, interventions worked only for individuals with low implicit bias scores, and

reversed for those with higher scores, raising concerns about unintended consequences (Anderson et al., 2015). These findings reflect the complexity of implicit bias as a construct and the difficulty of targeting it directly without clarity on how it operates in decision-making. This points to the broader limitation of individual-level interventions being rarely precision-targeted and often resting on assumptions about mechanisms (e.g., that association change leads to behaviour change) that may not hold in practice (Forscher et al., 2019). These approaches can still have value, but they need to be used carefully and only where there is clarity on what they target and how that connects to better decisions. Nonetheless, these results reinforce the view that targeting implicit bias through attitude change alone is unlikely to produce robust decision improvements.

In terms of transferability, the pattern is also clear and favours systemic interventions. What embeds most easily into workflow is also likely to scale well across contexts and roles. As such, changes to forms, processes, and displays are both practical to implement, typically requiring no facilitation, minimal costs and time burden, and highly transferable, particularly where decisions are repeated and structure can be applied without disrupting professional judgement. However, these same conditions that make these approaches transferable in principle may also constrain their implementation in practice. Specifically, in high-stakes domains such as forensic and legal work, professionals most affected by bias-related outcomes often operate under heavy time pressures, limited resources, and strict accountability demands. These realities can restrict both the feasibility and willingness to adopt new interventions, particularly where their potential influence on decision outcomes remains uncertain or where practitioners may later be required to defend their methods in court. Recognising these constraints is essential for interpreting how intervention findings translate beyond experimental settings, and for understanding why even well-evidenced strategies may prove difficult to embed within systems already functioning under significant procedural and ethical demands. Therefore, as transferability is shaped by both intervention design and the conditions under which they must operate, interventions that align with existing workflows and accountability structures have the highest potential to be sustained once implemented.

By contrast, resource-intensive strategies, such as high-fidelity simulations, while effective, remain more difficult to scale. However, their underlying logic, such as skill rehearsal and feedback, could still be approximated through lower-cost options, including video-supported reflective practice or digitally delivered simulations; though, such adaptation would need careful piloting to ensure they still deliver the same effects. These differences in practicality and transferability matter for high-stakes environments like forensic and legal contexts. In such settings, even modest improvements in decision quality could have significant downstream effects. Thus, systemic strategies that can be more readily implemented into existing processes offer a scalable route to improving consistency and fairness.

5. Conclusions

To advance from promising ideas to applied solutions, more research needs to test interventions under realistic constraints. Most studies in this review used simplified tasks, and mock professionals or adults with no professional relevance. Ecological realism remained limited across much of the evidence base, with only two studies (5.3%) demonstrating strong effects in high-stakes applied settings (Hamm et al., 2020; James et al., 2023). This limited what could be inferred about transfer to real-world contexts. To address this important gap, future applied testing should involve professional decision-makers, more realistic decision formats and tasks, stakeholder-informed designs and measurements that capture error, fidelity, and suitability for the intended setting. Systemic interventions, in

particular, should be evaluated not just for effect but for how they are used and translated in practice, including whether criteria and protocols are closely followed, and how reliably decision outcomes align with the intended standards.

Nevertheless, the applied potential of systemic strategies remains high, particularly for professional settings where decisions follow a predictable structure and are time-pressured and subject to wide discretion, such as forensic and legal contexts. The results reflect a broader shift in the field, after years of focus on reducing IAT scores that have yielded little practical value (Paluck et al., 2021), supporting current recommendations to instead target the context in which biased decisions are likely to occur, even before they occur (Axt et al., 2025; Greenwald et al., 2022). Where bias operates through fast, intuitive judgements, the most reliable strategy is to limit discretion and tie choices to predefined, evidence-based standards. Systemic strategies deliver the most consistent improvements and are typically easier to implement and scale. However, individual strategies add value when they teach the specific cognitive skills a decision might require in realistic simulated settings. For forensic, legal, and other high-stakes systems, where the consequences of error are serious, this combined approach, using systemic design to constrain bias-prone moments and targeted individual training to support improved judgement under pressure, offers the most encouraging route to fairer outcomes.

Beyond the decisions of judges, jurors and other judicial decision-makers, the forensic domain more broadly also warrants attention. Judicial outcomes are often shaped by the evaluations of forensic experts, whose assessments can carry considerable weight in legal reasoning. Emerging evidence indicates that such evaluations are themselves vulnerable to implicit and cognitive biases (Buongiorno et al., 2025), raising the possibility of a cascading effect, in which biases introduced during expert assessment may continue to shape judicial interpretation and reasoning. Our review supports the view that empirical work in this area remains limited, and future research should extend applied testing to forensic expert contexts to examine whether bias-reduction interventions can operate effectively at these earlier stages of decision-making, interrupting potential ‘cascades’ before they reach judicial outcomes. Such work would help clarify how systemic and individual-level strategies can be adapted to strengthen fairness and consistency throughout the justice process.

More broadly, addressing these interconnected sources of bias is essential for translating promising interventions into applied practice and ensuring their impact across professional domains. However, while more applied testing is needed to strengthen the evidence base, especially under realistic constraints, the current evidence base suggests that effective solutions might already exist, but they are yet to be embedded and evaluated in the settings where they are most needed.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/bs15111592/s1>. Full database-specific search strings, fields and filters are presented as a Supplementary Material (S1), reproducing the exact configurations used across all eight databases. A Searchable Evidence Map (interactive, filterable dataset of study and intervention characteristics and outcomes) is also available open-access as supplementary material via OSF at <https://doi.org/10.17605/OSF.IO/NE4DV>.

Author Contributions: Conceptualization, I.M., F.G. and A.J.S.; methodology, I.M., F.G. and A.J.S.; software, I.M.; validation, I.M., F.G. and A.J.S.; formal analysis, I.M.; investigation, I.M.; resources, I.M., F.G. and A.J.S.; data curation, I.M.; writing—original draft preparation, I.M.; writing—review and editing, I.M., F.G. and A.J.S.; visualization, I.M.; supervision, F.G. and A.J.S.; project administration, I.M.; funding acquisition, I.M., F.G. and A.J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Economic and Social Research Council (ESRC) through the South and East Network for Social Sciences (SeNSS) Doctoral Training Partnership [grant number ES/Y001834/1].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: A Searchable Evidence Map (interactive, filterable dataset of study and intervention characteristics and outcomes) is available open-access as Supplementary Material via OSF <https://doi.org/10.17605/OSF.IO/NE4DV>. The full Searchable Systematic Map with extracted dataset, coding materials, and working notes is available from the corresponding author on reasonable request. All underlying sources are published studies cited in the article; no new primary data were generated.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Full PsycInfo Search String

The adapted strings, search fields, and database-supported filters for each of the eight databases and platforms are provided in Supplementary Material (S1).

("implicit" OR "unconscious" OR "subconscious" OR "automatic" OR "heuristic" OR "myth*") AND ("bias*" OR "prejudic*" OR "stereotyp*" OR "attitude*" OR "association*" OR "discriminat*" OR "preference*" OR "myth*") AND ("debias*" OR "intervention*" OR "reduc*" OR "training" OR "strategy" OR "method" OR "program*" OR "approach*" OR "chang*" OR "tool*" OR "educat*" OR "modif*" OR "diminish*" OR "counteract*" OR "mitigat*" OR "refram*" OR "effective*") AND ("decision-making" OR "decision task" OR "judg*" OR "evaluat*" OR "deliberation*" OR "verdict")

Appendix B

Data Extraction Framework and Coding Domains

A structured narrative synthesis was conducted to examine whether and how included interventions produced reductions in bias, and to identify patterns across intervention types, decision contexts, and outcome characteristics. The synthesis was guided by four domains used in the SSM: 1. Study details, 2. Intervention characteristics, 3. Outcomes and findings, and 4. Delivery practicality and implementation feasibility. These domains and their dimensions are described below.

1. Study details. This domain captured core methodological and contextual features of each study, including year of publication, country, professional field, study design, participant role, and paradigm. It also included three indicators of ecological realism, which were used to assess the ecological validity of each study and determine whether the intervention was tested under highly controlled or realistically applied conditions.

Given the review's aim to identify interventions that are not only effective but also practical and transferable, it is essential to consider the extent to which intervention effects are likely to hold under real-world conditions, especially in high-stakes settings, where time, cognitive resources and institutional constraints may differ substantially from laboratory or student-based studies.

The ecological indicators included in the SSM are as follows. The first indicator is sample realism, which captured whether the participants were actual professionals making decisions in their own field (high realism), trainees with professional relevance (moderate realism), or general student samples without contextual relevance (low realism). The second indicator is task realism, which refers to the nature of the evaluative decision itself, and whether it directly reflects a professional decision with practical consequences (high

realism), simulates a real decision but in a simplified or lower-stakes format (moderate realism), or involves abstract or artificial tasks (low realism). The third indicator is context realism, which was coded as an integrated measure that considered the alignment of sample, task, and setting, reflecting the overall fidelity of the study's design to applied professional conditions. These three ecological indicators will be fused to evaluate the generalisability of findings to high-stakes legal and related professional decision-making contexts.

2. Intervention characteristics. This domain captured core features of how each intervention was designed, delivered, and intended to work. It included the intervention's mechanism of action, delivery method and materials, timing relative to the decision task and overall duration. To support structured analysis and cross-study comparison, two classifications were applied to each intervention as follows.

The first classification coded interventions according to the level at which they operated, reflecting whether the strategy targeted the individual decision-maker, the structure of the decision context, or both. This approach builds on a growing distinction in the literature between interventions that seek to change internal cognitive or motivational processes and those that aim to modify the external conditions under which decisions are made (Axt et al., 2025; Greenwald et al., 2022). Interventions were therefore coded as individual-level, systemic-level, or mixed. Individual-level interventions aimed to influence cognitive, affective or motivational processes, often by prompting the decision-maker to consciously reflect on, interrupt, or override biased tendencies. These included strategies such as stereotype replacement, mindfulness, reflective prompts, and exposure to counter-stereotypical exemplars. Systemic-level interventions modified the structure of the decision environment to reduce the influence of bias regardless of the decision-maker's intent or awareness. These included strategies such as blind evaluations, structured interviews, or scoring rubrics, which aimed to constrain discretion, standardise criteria, or remove identifying cues. Mixed interventions integrated both approaches, combining cognitive or motivational strategies with structural supports, for example, embedding self-regulation prompts into structured forms, or pairing training with procedural changes that guide or reinforce equitable decision-making.

The second classification grouped interventions according to the specific mechanism through which they aimed to reduce bias in decision-making. These categories were developed to reflect what each intervention changed, disrupted, or introduced into the decision process, and were designed to capture key distinctions in how interventions function, supporting more fine-grained analysis of intervention strategy and design. Interventions were coded as: changing what decision-makers see at the point of judgement, adding structure that limits discretion, prompting self-regulation, reframing beliefs or assumptions, and targeting automatic associations. This classification will be used in the synthesis to compare interventions by mechanism, enabling the review to identify whether certain types of strategies tend to produce more consistent, durable or transferable effects across different decision settings.

3. Outcomes and findings. This domain captured the results reported in each study, with a focus on whether the intervention produced a measurable reduction in biased outcomes. Effectiveness of intervention was defined as a meaningful change in the evaluative decision the intervention aimed to influence, conditional on evidence of baseline bias. Baseline bias was defined as a measurable disparity or pattern of biased decision-making prior to the intervention, either within the intervention group or demonstrated through a control or comparison group. In the absence of a pre-intervention bias, no claims were made about effectiveness, as any observed changes could not be attributed to the intervention as evidence of bias reduction.

To assess effectiveness, outcomes were categorised into two groups: targeted outcomes and other observed outcomes. Targeted outcomes were defined as the specific evaluative decisions the intervention was designed to influence and were the primary basis for assessing effectiveness. These included applied judgements such as hiring decisions, sentencing outcomes, performance ratings, or treatment recommendations, where bias was evidenced by differences in how individuals were evaluated based on characteristics such as race, gender, or age. A reduction in bias was defined as a statistically meaningful shift in those evaluative patterns following the intervention. Other observed outcomes referred to additional effects that were not the primary focus of the intervention and were not used to determine effectiveness. These offered relevant insight into the broader impact of the intervention, including aspects such as changes in implicit or explicit bias measures (e.g., IAT scores), secondary decision variables, or qualitative data on participant experience.

Each targeted outcome was appraised for strength of evidence, based on whether effects were statistically significant, consistent, and directionally aligned with the intervention's aims. Outcomes were then summarised according to the intervention's impact on both the primary target and other observed variables, allowing structured comparison across decision types and intervention strategies. Where available, data on the durability of effects were also extracted. This referred to whether the intervention's impact persisted beyond the immediate task, for example, through delayed post-tests, longitudinal follow-up or repeated decision opportunities. This information was used to appraise the temporal stability of intervention effects, which is an important consideration for real-world applicability.

4. Delivery practicality and implementation feasibility. This domain captured the conditions required to deliver each intervention in real settings. For each study, format, materials, time demands, and the level of expertise or training required to implement the intervention effectively were recorded. Two constructs were central to interpretation and coding in this domain: delivery practicality and implementation feasibility. These definitions were used consistently across studies to relate evidence of effectiveness to the conditions under which interventions are delivered and maintained in real settings.

Delivery practicality was defined as the concrete resource and process requirements needed to use an intervention as designed. This included both direct and indirect costs, such as the need for equipment and materials, the time burden placed on decision-makers, the level of training and supervision required for facilitators, and the extent to which the intervention aligned with existing workflows, systems, and policies. This construct addressed whether a typical organisation could adopt the intervention using the resources it already had or could reasonably acquire.

Implementation feasibility referred to the likelihood that an intervention would be adopted, delivered with sufficient fidelity, and sustained over time in the target setting and in comparable environments. It focused on how well an intervention would perform under real-world constraints, including time pressure, staff variability, governance requirements, and the availability of training or oversight. This construct assessed how readily an intervention could be delivered in its tested setting, and how easily its core mechanism could be applied in other person-evaluation contexts with minimal adaptation.

To structure this domain, six delivery-related aspects were assessed for each study. These were based on reported information and, where needed, supplemented by reasonable inferences about practical demands. As the review spans multiple applied settings, including education, healthcare, and workplace settings, judgements regarding scalability were informed by generalised reasoning rather than field-specific expertise and should be interpreted accordingly.

First, cost practicality captured the intervention’s resource needs, based on materials, equipment, personnel and any operational costs. Second, facilitation practicality assessed whether delivery could be carried out independently by participants or whether it required trained facilitators or supervision. Third, duration practicality referred to the time burden associated with the intervention, including session length and scheduling challenges. Fourth, practical scalability in the target field examined how easily the intervention could be rolled out more widely within the setting in which it was tested, considering compatibility with existing workflows, policies, and delivery systems. Fifth, the field scalability of the intervention approach considered how transferable the underlying mechanism was to other decision-making domains with only minimal adaptations. Lastly, overall deployability provided an integrated judgement of whether the intervention was likely to be adopted and sustained under typical organisational constraints, reflecting a synthesis of the five dimensions above. Ratings of high, moderate, or low were applied to each dimension. High ratings indicated minimal resource needs, low delivery burden, and strong compatibility with existing practice. Moderate ratings reflected delivery conditions that were feasible but involved certain limitations, added steps, or resource considerations that could affect ease of implementation. Where detail was insufficient to support a firm judgement, ratings were conservatively assigned and flagged as assumed from context.

Appendix C

QuADS Scoring Sheet

This table presents the full set of QuADS domains and their corresponding 0–3 scoring anchors, as used in the quality appraisal of included studies. The QuADS framework provides a structured approach for assessing methodological quality and reporting transparency across diverse study designs.

Criterion	0	1	2	3
1. Theoretical or conceptual underpinning to the research	No mention at all.	General reference to broad theories or concepts that frame the study. e.g., key concepts were identified in the introduction section.	Identification of specific theories or concepts that frame the study and how these informed the work undertaken. e.g., key concepts were identified in the introduction section and applied to the study.	Explicit discussion of the theories or concepts that inform the study, with application of the theory or concept evident through the design, materials and outcomes explored. e.g., key concepts were identified in the introduction section and the application apparent in each element of the study design.
2. Statement of research aim/s	No mention at all.	Reference to what the study sought to achieve embedded within the report but no explicit aims statement.	Aims statement made but may only appear in the abstract or be lacking detail.	Explicit and detailed statement of aim/s in the main body of report.

3. Clear description of research setting and target population	No mention at all.	General description of research area but not of the specific research environment e.g., 'in primary care.'	Description of research setting is made but is lacking detail e.g., 'in primary care practices in region [x]'.	Specific description of the research setting and target population of study e.g., 'nurses and doctors from GP practices in [x] part of [x] city in [x] country.'
4. The study design is appropriate to address the stated research aim/s	No research aim/s stated or the design is entirely unsuitable e.g., a Y/N item survey for a study seeking to undertake exploratory work of lived experiences.	The study design can only address some aspects of the stated research aim/s e.g., use of focus groups to capture data regarding the frequency and experience of a disease.	The study design can address the stated research aim/s but there is a more suitable alternative that could have been used or used in addition e.g., addition of a qualitative or quantitative component could strengthen the design.	The study design selected appears to be the most suitable approach to attempt to answer the stated research aim/s.
5. Appropriate sampling to address the research aim/s	No mention of the sampling approach.	Evidence of consideration of the sample required e.g., the sample characteristics are described and appear appropriate to address the research aim/s.	Evidence of consideration of sample required to address the aim. e.g., the sample characteristics are described with reference to the aim/s.	Detailed evidence of consideration of the sample required to address the research aim/s. e.g., sample size calculation or discussion of an iterative sampling process with reference to the research aims or the case selected for study.
6. Rationale for choice of data collection tool/s	No mention of rationale for data collection tool used.	Very limited explanation for choice of data collection tool/s. e.g., based on availability of tool.	Basic explanation of rationale for choice of data collection tool/s. e.g., based on use in a prior similar study.	Detailed explanation of rationale for choice of data collection tool/s. e.g., relevance to the study aim/s, co-designed with the target population or assessments of tool quality.

7. The format and content of data collection tool is appropriate to address the stated research aim/s	No research aim/s stated and/or data collection tool not detailed.	Structure and/or content of tool/s suitable to address some aspects of the research aim/s or to address the aim/s superficially e.g., single item response that is very general or an open-response item to capture content which requires probing.	Structure and/or content of tool/s allow for data to be gathered broadly addressing the stated aim/s but could benefit from refinement. e.g., the framing of survey or interview questions are too broad or focused to one element of the research aim/s.	Structure and content of tool/s allow for detailed data to be gathered around all relevant issues required to address the stated research aim/s.
8. Description of data collection procedure	No mention of the data collection procedure.	Basic and brief outline of data collection procedure e.g., ‘using a questionnaire distributed to staff’.	States each stage of data collection procedure but with limited detail or states some stages in detail but omits others e.g., the recruitment process is mentioned but lacks important details.	Detailed description of each stage of the data collection procedure, including when, where and how data was gathered, such that the procedure could be replicated.
9. Recruitment data provided	No mention of recruitment data.	Minimal and basic recruitment data e.g., number of people invited who agreed to take part.	Some recruitment data but not a complete account e.g., number of people who were invited and agreed.	Complete data allowing for full picture of recruitment outcomes e.g., number of people approached, recruited, and who completed with attrition data explained where relevant.
10. Justification for analytic method selected	No mention of the rationale for the analytic method chosen.	Very limited justification for choice of analytic method selected. e.g., previous use by the research team.	Basic justification for choice of analytic method selected e.g., method used in prior similar research.	Detailed justification for choice of analytic method selected e.g., relevance to the study aim/s or comment around the strengths of the method selected.

11. The method of analysis was appropriate to answer the research aim/s	No mention at all.	Method of analysis can only address the research aim/s basically or broadly.	Method of analysis can address the research aim/s, but there is a more suitable alternative that could have been used or used in addition to offer a stronger analysis.	Method of analysis selected is the most suitable approach to attempt answer the research aim/s in detail e.g., for qualitative interpretative phenomenological analysis might be considered preferable for experiences vs. content analysis to elicit frequency of occurrence of events.
12. Evidence that the research stakeholders have been considered in the research design or conduct.	No mention at all.	Consideration of some of the research stakeholders e.g., use of pilot study with target sample but no stakeholder involvement in planning stages of study design.	Evidence of stakeholder input informing the research. e.g., use of a pilot study with feedback influencing the study design/conduct or reference to a project reference group established to guide the research.	Substantial consultation with stakeholders identifiable in planning of study design and in preliminary work e.g., consultation in the conceptualisation of the research, a project advisory group or evidence of stakeholder input informing the work.
13. Strengths and limitations critically discussed	No mention at all.	Very limited mention of strengths and limitations, with omissions of many key issues. e.g., one or two strengths/limitations mentioned with limited detail.	Discussion of some of the key strengths and weaknesses of the study, but not complete. e.g., several strengths/limitations explored but with notable omissions or lack of depth of explanation.	Thorough discussion of strengths and limitations of all aspects of study including design, methods, data collection tools, sample & analytic approach.

References

Note: References marked with an asterisk (*) were included in the systematic review; unmarked references are cited for background or discussion.

- * Anderson, A. J., Ahmad, A. S., King, E. B., Lindsey, A. P., Feyre, R. P., Ragone, S., & Kim, S. (2015). The effectiveness of three strategies to reduce the influence of bias in evaluations of female leaders. *Journal of Applied Social Psychology*, 45(9), 522–539. [\[CrossRef\]](#)
- Axt, J., Posada, V. P., Roy, E., & To, J. (2025). Revisiting the policy implications of implicit social cognition. *Social Issues and Policy Review*, 19(1), e70003. [\[CrossRef\]](#)

- * Bragger, J. D., Kutcher, E., Morgan, J., & Firth, P. (2002). The effects of the structured interview on reducing biases against pregnant job applicants. *Sex Roles*, 46(7), 215–226. [\[CrossRef\]](#)
- * Brauer, M., & Er-rafiy, A. (2011). Increasing perceived variability reduces prejudice and discrimination. *Journal of Experimental Social Psychology*, 47(5), 871–881. [\[CrossRef\]](#)
- Buongiorno, L., Mele, F., Petroni, G., Margari, A., Carabellese, F., Catanesi, R., & Mandarelli, G. (2025). Cognitive biases in forensic psychiatry: A scoping review. *International Journal of Law and Psychiatry*, 101, 102083. [\[CrossRef\]](#)
- Buttrick, N., Axt, J., Ebersole, C. R., & Huband, J. (2020). Re-assessing the incremental predictive validity of Implicit Association Tests. *Journal of Experimental Social Psychology*, 88, 103941. [\[CrossRef\]](#)
- Curley, L. J., Munro, J., & Dror, I. E. (2022). Cognitive and human factors in legal layperson decision making: Sources of bias in juror decision making. *Medicine, Science and the Law*, 62(3), 206–215. [\[CrossRef\]](#) [\[PubMed\]](#)
- Curley, L. J., Munro, J., Lages, M., MacLean, R., & Murray, J. (2020). Assessing cognitive bias in forensic decisions: A review and outlook. *Journal of Forensic Sciences*, 65(2), 354–360. [\[CrossRef\]](#)
- Curley, L. J., & Neuhaus, T. (2024). Are legal experts better decision makers than jurors? A psychological evaluation of the role of juries in the 21st century. *Journal of Criminal Psychology*, 14(4), 325–335. [\[CrossRef\]](#)
- * Dahlen, B., McGraw, R., & Vora, S. (2024). Evaluation of simulation-based intervention for implicit bias mitigation: A response to systemic racism. *Clinical Simulation in Nursing*, 95, 101596. [\[CrossRef\]](#)
- De Houwer, J. (2019). Implicit bias is behavior: A functional-cognitive perspective on implicit bias. *Perspectives on Psychological Science*, 14(5), 835–840. [\[CrossRef\]](#)
- * Derous, E., Nguyen, H.-H. D., & Ryan, A. M. (2021). Reducing ethnic discrimination in resume-screening: A test of two training interventions. *European Journal of Work and Organizational Psychology*, 30(2), 225–239. [\[CrossRef\]](#)
- Dobbie, W., Goldin, J., & Yang, C. S. (2018). The effects of pre-trial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *American Economic Review*, 108(2), 201–240. [\[CrossRef\]](#)
- * Döbrich, C., Wollersheim, J., Welp, I. M., & Spörrle, M. (2014). Debiasing age discrimination in HR decisions. *International Journal of Human Resources Development and Management*, 14(4), 219. [\[CrossRef\]](#)
- Dror, I. E. (2025). Biased and biasing: The hidden bias cascade and bias snowball effects. *Behavioral Sciences*, 15(4), 490. [\[CrossRef\]](#) [\[PubMed\]](#)
- Dror, I. E., Melinek, J., Arden, J. L., Kukucka, J., Hawkins, S., Carter, J., & Atherton, D. S. (2021). Cognitive bias in forensic pathology decisions. *Journal of Forensic Sciences*, 66(5), 1751–1757. [\[CrossRef\]](#)
- Edkins, V. A. (2011). Defense attorney plea recommendations and client race: Does zealous representation apply equally to all? *Law and Human Behavior*, 35(5), 413–425. [\[CrossRef\]](#) [\[PubMed\]](#)
- * Feng, Z., Liu, Y., Wang, Z., & Savani, K. (2020). Let's choose one of each: Using the partition dependence effect to increase diversity in organizations. *Organizational Behavior and Human Decision Processes*, 158, 11–26. [\[CrossRef\]](#)
- FitzGerald, C., & Hurst, S. (2017). Implicit bias in healthcare professionals: A systematic review. *BMC Medical Ethics*, 18(1), 19. [\[CrossRef\]](#)
- FitzGerald, C., Martin, A., Berner, D., & Hurst, S. (2019). Interventions designed to reduce implicit prejudices and implicit stereotypes in real world contexts: A systematic review. *BMC Psychology*, 7(1), 29. [\[CrossRef\]](#)
- Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology*, 117(3), 522–559. [\[CrossRef\]](#)
- * Friedmann, E., & Efrat-Treister, D. (2023). Gender bias in stem hiring: Implicit in-group gender favoritism among men managers. *Gender & Society*, 37(1), 32–64. [\[CrossRef\]](#)
- Gawronski, B., Ledgerwood, A., & Eastwick, P. W. (2022). Implicit bias $\neq$ bias on implicit measures. *Psychological Inquiry*, 33(3), 139–155. [\[CrossRef\]](#)
- Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102(1), 4–27. [\[CrossRef\]](#)
- Greenwald, A. G., Dasgupta, N., Dovidio, J. F., Kang, J., Moss-Racusin, C. A., & Teachman, B. A. (2022). Implicit-bias remedies: Treating discriminatory bias as a public-health problem. *Psychological Science in the Public Interest*, 23(1), 7–40. [\[CrossRef\]](#) [\[PubMed\]](#)
- * Hamm, R. F., Srinivas, S. K., & Levine, L. D. (2020). A standardized labor induction protocol: Impact on racial disparities in obstetrical outcomes. *American Journal of Obstetrics & Gynecology MFM*, 2(3), 100148. [\[CrossRef\]](#)
- Harrison, R., Jones, B., Gardner, P., & Lawton, R. (2021). Quality assessment with diverse studies (QuADS): An appraisal tool for methodological and reporting quality in systematic reviews of mixed- or multi-method studies. *BMC Health Services Research*, 21(1), 144. [\[CrossRef\]](#)
- * Hirsh, A. T., Miller, M. M., Hollingshead, N. A., Anastas, T., Carnell, S. T., Lok, B. C., Chu, C., Zhang, Y., Robinson, M. E., Kroenke, K., & Ashburn-Nardo, L. (2019). A randomized controlled trial testing a virtual perspective-taking intervention to reduce race and socioeconomic status disparities in pain care. *Pain*, 160(10), 2229–2240. [\[CrossRef\]](#)

- Holroyd, J., & Sweetman, J. (2016). The heterogeneity of implicit bias. In M. Brownstein, & J. Saul (Eds.), *Implicit bias and philosophy* (Vol. 1, pp. 80–103). Oxford University Press. [CrossRef]
- Hopkins, K., Uhrig, N., & Colahan, M. (2016). *Associations between ethnic background and being sentenced to prison in the Crown Court in England and Wales in 2015*. Ministry of Justice. Available online: <https://assets.publishing.service.gov.uk/media/5a814a3d40f0b62305b8e241/associations-between-ethnic-background-being-sentenced-to-prison-in-the-crown-court-in-england-and-wales-2015.pdf> (accessed on 19 August 2025).
- * James, L., James, S., & Mitchell, R. J. (2023). Results from an effectiveness evaluation of anti-bias training on police behavior and public perceptions of discrimination. *Policing: An International Journal*, 46(5/6), 831–845. [CrossRef]
- * Kawakami, K., Dovidio, J. F., & Van Kamp, S. (2005). Kicking the habit: Effects of nonstereotypic association training and correction processes on hiring decisions. *Journal of Experimental Social Psychology*, 41(1), 68–75. [CrossRef]
- * Kawakami, K., Dovidio, J. F., & Van Kamp, S. (2007). The impact of counterstereotypic training and related correction processes on the application of stereotypes. *Group Processes & Intergroup Relations*, 10(2), 139–156. [CrossRef]
- * Kleissner, V., & Jahn, G. (2021). Implicit and explicit age cues influence the evaluation of job applications. *Journal of Applied Social Psychology*, 51(2), 107–120. [CrossRef]
- Kovera, M. B. (2019). Racial disparities in the criminal justice system: Prevalence, causes, and a search for solutions. *Journal of Social Issues*, 75(4), 1139–1164. [CrossRef]
- Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G., & Banaji, M. R. (2019). Relationship between the implicit association test and intergroup behavior: A meta-analysis. *American Psychologist*, 74(5), 569–586. [CrossRef]
- Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., Calanchini, J., Xiao, Y. J., Pedram, C., Marshburn, C. K., Simon, S., Blanchar, J. C., Joy-Gaba, J. A., Conway, J., Redford, L., Klein, R. A., Roussos, G., Schellhaas, F. M. H., Burns, M., . . . Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, 145(8), 1001–1016. [CrossRef]
- * Lehmann-Grube, S. K., Tobisch, A., & Dresel, M. (2024). Changing preservice teacher students' stereotypes and attitudes and reducing judgment biases concerning students of different family backgrounds: Effects of a short intervention. *Social Psychology of Education*, 27(4), 1621–1658. [CrossRef]
- * Liu, Z., Rattan, A., & Savani, K. (2023). Reducing gender bias in the evaluation and selection of future leaders: The role of decision-makers' mindsets about the universality of leadership potential. *Journal of Applied Psychology*, 108(12), 1924–1951. [CrossRef] [PubMed]
- * Lucas, B. J., Berry, Z., Giurge, L. M., & Chugh, D. (2021). A longer shortlist increases the consideration of female candidates in male-dominant domains. *Nature Human Behaviour*, 5(6), 736–742. [CrossRef]
- * Lynch, M., Kidd, T., & Shaw, E. (2022). The subtle effects of implicit bias instructions. *Law & Policy*, 44(1), 98–124. [CrossRef]
- Mitchell, T. L., Haw, R. M., Pfeifer, J. E., & Meissner, C. A. (2005). Racial bias in mock juror decision-making: A meta-analytic review of defendant treatment. *Law and Human Behavior*, 29(6), 621–637. [CrossRef]
- Mustard, D. B. (2001). Racial, ethnic, and gender disparities in sentencing: Evidence from the U.S. federal courts. *The Journal of Law and Economics*, 44(1), 285–314. [CrossRef]
- * Naser, S. C., Brann, K. L., & Noltemeyer, A. (2021). A brief report on the promise of system 2 cues for impacting teacher decision-making in school discipline practices for Black male youth. *School Psychology*, 36(3), 196–202. [CrossRef]
- * Neal, D., Morgan, J. L., Ormerod, T., & Reed, M. W. R. (2024). Intervention to reduce age bias in medical students' decision making for the treatment of older women with breast cancer: A novel approach to bias training. *Journal of Psychosocial Oncology*, 42(1), 48–63. [CrossRef] [PubMed]
- Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, 105(2), 171–192. [CrossRef]
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. [CrossRef]
- Paluck, E. L., Porat, R., Clark, C. S., & Green, D. P. (2021). Prejudice reduction: Progress and challenges. *Annual Review of Psychology*, 72(1), 533–560. [CrossRef]
- * Pershing, S., Stell, L., Fisher, A. C., & Goldberg, J. L. (2021). Implicit bias and the association of redaction of identifiers with residency application screening scores. *JAMA Ophthalmology*, 139(12), 1274. [CrossRef] [PubMed]
- Pfeifer, J. E., & Ogloff, J. R. P. (1991). Ambiguity and guilt determinations: A modern racism perspective1. *Journal of Applied Social Psychology*, 21(21), 1713–1725. [CrossRef]
- * Quinn, D. M. (2020). Experimental evidence on teachers' racial bias in student evaluation: The role of grading scales. *Educational Evaluation and Policy Analysis*, 42(3), 375–392. [CrossRef]

- Rehavi, M. M., & Starr, S. B. (2014). Racial disparity in federal criminal sentences. *Journal of Political Economy*, 122(6), 1320–1354. [\[CrossRef\]](#)
- * Ruva, C. L., Sykes, E. C., Smith, K. D., Deaton, L. R., Erdem, S., & Jones, A. M. (2024). Battling bias: Can two implicit bias remedies reduce juror racial bias? *Psychology, Crime & Law*, 30(7), 730–757. [\[CrossRef\]](#)
- Sah, S., Robertson, C. T., & Baughman, S. B. (2015). Blinding prosecutors to defendants' race: A policy proposal to reduce unconscious bias in the criminal justice system. *Behavioral Science & Policy*, 1(2), 69–76. [\[CrossRef\]](#)
- Sah, S., Tannenbaum, D., Cleary, H., Feldman, Y., Glaser, J., Lerman, A., MacCoun, R., Maguire, E., Slovic, P., Spellman, B., Spohn, C., & Winship, C. (2016). Combating biased decisionmaking & promoting justice & equal treatment. *Behavioral Science & Policy*, 2(2), 79–87. [\[CrossRef\]](#) [\[PubMed\]](#)
- * Salmanowitz, N. (2018). The impact of virtual reality on implicit racial bias and mock legal decisions. *Journal of Law and the Biosciences*, 5(1), 174–203. [\[CrossRef\]](#)
- Sargent, M. J., & Bradfield, A. L. (2004). Race and information processing in criminal trials: Does the defendant's race affect how the facts are evaluated? *Personality and Social Psychology Bulletin*, 30(8), 995–1008. [\[CrossRef\]](#)
- Schlesinger, T. (2005). Racial and ethnic disparity in pretrial criminal processing. *Justice Quarterly*, 22(2), 170–192. [\[CrossRef\]](#)
- Sommers, S. R., & Ellsworth, P. C. (2001). White juror bias: An investigation of prejudice against black defendants in the American courtroom. *Psychology, Public Policy, and Law*, 7(1), 201–229. [\[CrossRef\]](#)
- * Uhlmann, E. L., & Cohen, G. L. (2005). Constructed criteria: Redefining merit to justify discrimination. *Psychological Science*, 16(6), 474–480. [\[CrossRef\]](#) [\[PubMed\]](#)
- Vela, M. B., Erondur, A. I., Smith, N. A., Peek, M. E., Woodruff, J. N., & Chin, M. H. (2022). Eliminating explicit and implicit biases in health care: Evidence and research needs. *Annual Review of Public Health*, 43(1), 477–501. [\[CrossRef\]](#) [\[PubMed\]](#)
- * Wall, E., Narechania, A., Coscia, A., Paden, J., & Endert, A. (2022). Left, right, and gender: Exploring interaction traces to mitigate human biases. *IEEE Transactions on Visualization and Computer Graphics*, 28(1), 966–975. [\[CrossRef\]](#)
- Young, D. M., Levinson, J. D., & Sinnett, S. (2014). Innocent until primed: Mock jurors' racially biased response to the presumption of innocence. *PLoS ONE*, 9(3), e92365. [\[CrossRef\]](#) [\[PubMed\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.