Köln Z Soziol
https://doi.org/10.1007/s11577-026-01071-y

Kölner Zeitschrift für Soziologie und Sozialpsychologie

RESEARCH ARTICLE

Text as Data and Causal Inference in Sociology

Nicole Schwitter · Ruben L. Bach · Christopher Klamm

Received: 6 October 2025 / Accepted: 24 March 2026

Abstract This article discusses and showcases approaches to causal inference with text as data, focusing on the challenges and opportunities that arise when sociological constructs are embedded in language. We highlight that text embeds latent constructs and comes with a number of measurement challenges: Variables of interest must be interpreted through coding, feature extraction, and modelling choices. We distinguish between three central designs: when text is the outcome to be explained (text as outcome), when text functions as the treatment whose effects are to be estimated (text as treatment), and when information extracted from text is used as control variable (text as control). Drawing on examples from our own research, we illustrate how recent methodological advances allow researchers to learn latent outcomes or latent treatments from textual data while preserving the logic of causal identification. We use state-of-the-art techniques for drawing causal inferences with text as data and show how text can serve as a window into sociologically relevant constructs, while also underscoring the interpretive leeway inherent in computational modelling. Our analysis demonstrates that causal inference with text requires careful attention to theory, transparency about researchers' choices, and sensitivity to the data-generating

✉ N. Schwitter · R. L. Bach
Mannheim Centre for European Social Research, University of Mannheim
Mannheim, Germany
E-Mail: Nicole.schwitter@uni-mannheim.de

✉ N. Schwitter
Department of Sociology, University of Warwick
Coventry, UK

C. Klamm
School of Business Informatics and Mathematics, University of Mannheim
Mannheim, Germany

University of Cologne
Cologne, Germany

Published online: 23 April 2026

process. We conclude that text-as-data approaches hold promise for causal analysis and can contribute to sociological explanation.

Keywords Natural Language Processing (NLP) · Survey experiments · Measurement · Latent treatment · Criminality · Gender inequality

Textdaten und kausale Inferenz in der Soziologie

Zusammenfassung Dieser Artikel diskutiert und zeigt Ansätze zur kausalen Inferenz mit Textdaten auf. Wir gehen im Artikel auf die Herausforderungen und Chancen ein, die entstehen, wenn soziologische Konstrukte in Sprache eingebettet sind. Wir betonen, dass Texte latente Konstrukte enthalten und mit verschiedenen Messproblemen einhergehen: Die interessierenden Variablen müssen über Kodierung, Feature-Extraktion und Modellierungsentscheidungen interpretiert werden. Wir unterscheiden zwischen drei zentralen Designs: erstens, wenn Text das zu erklärende Outcome ist (*text as outcome*), zweitens, wenn Text als Treatment fungiert, dessen Effekte geschätzt werden sollen (*text as treatment*) und drittens, wenn Informationen aus Texten als Kontrollvariablen genutzt werden (*text as control*). Anhand von Beispielen aus unserer eigenen Forschung veranschaulichen wir, wie aktuelle methodische Fortschritte es ermöglichen, latente Outcomes oder Treatments aus Textdaten zu lernen, ohne die Logik kausaler Identifikation zu verletzen. In unseren Anwendungsfällen nutzen wir moderne Ansätze zur kausalen Inferenz mit Textdaten und zeigen, wie Texte als Fenster zu soziologisch relevanten Konstrukten dienen können, während sie gleichzeitig den interpretativen Spielraum in computergestützten Modellen verdeutlichen. Unsere Analyse zeigt, dass kausale Inferenz mit Text sorgfältige theoretische Überlegungen, Transparenz bei den Entscheidungen der Forschenden und Sensibilität für den datengenerierenden Prozess erfordert. Wir schließen, dass Text-als-Daten-Ansätze vielversprechend für kausale Analysen sind und zur soziologischen Erklärung beitragen können.

Schlüsselwörter Automatische Sprachverarbeitung · Surveyexperimente · Messung · Latentes Treatment · Kriminalität · Geschlechterungleichheit

1 Introduction

Textual data, ranging from open-ended survey responses and news articles to social media discourse, have become increasingly popular in the empirical social sciences (Hassan et al. 2025; Keuschnigg et al. 2018). This trend has largely been facilitated by advances in computational methods that allow large-scale processing of unstructured text (Benoit 2020; Grimmer et al. 2022; Grimmer and Stewart 2013). Yet although "text-as-data" approaches are gaining traction, their integration into sociological explanation—understood in the tradition of analytical sociology as the identification of micro-mechanisms that generate observed macro-sociological phenomena (Diekmann 2026, this issue)—and causal inference remains limited. Much

work with text as data in the social sciences is descriptive, focusing on classifying texts, detecting thematic patterns, or applying sentiment analysis (see, e.g., Bleich et al. 2015; Golder and Macy 2011). These approaches provide valuable empirical insights but often fall short of engaging with causal explanation and theory testing, which lie at the heart of sociological inquiry (for a discussion of theory testing with text data, see Hurtado Bodell et al. 2026, this issue). While sociological explanations ideally encompass both the identification of causal effects and the specification of the underlying mechanisms, in this paper we focus primarily on the methodological challenges of causal identification when using textual data. We emphasise that identifying a causal effect does not replace theoretical embedding or the modelling of social mechanisms; rather, it provides a necessary empirical foundation for them (Raub 2026, this issue). In line with the goals of this special issue, we treat causal identification as a component of, rather than a synonym for, sociological explanation.

Text data offer opportunities for causal inference. Because text is highly expressive, it can encode rich information about beliefs, motivations, reasoning, and social context that is difficult to capture using conventional variables alone. Moreover, text is often generated as a by-product of social processes and at scale (e.g. administrative records, online communication) rather than in response to research instruments and is therefore often less directly shaped by survey design or social desirability concerns. When carefully modelled, text-based representations can improve measurement, reduce omitted-variable bias, and provide empirical access to intermediate processes that are central to sociological explanation.

However, applying causal inference to textual data also comes with distinct challenges. Language and text are complex and, in a first step, need to be transformed into quantitative representations. In conventional causal inference designs in sociology, key variables—such as years of education, income, employment status, party identification, or experimentally assigned treatments—are typically well-defined prior to analysis and observed directly or with limited measurement uncertainty. In contrast, the key variables of interest in text-based studies are often latent, meaning that they must be inferred from the text itself. This complicates the application of standard causal inference frameworks, such as the potential outcomes model (Rubin 1974), which assume fixed, pre-specified outcomes and treatments.

Recent methodological contributions, particularly from computer science, computational social science, and political science (e.g. Egami et al. 2022; Fong and Grimmer 2023; Imai and Nakamura 2024, 2025; Yang and Shen 2025), have begun to address these issues. They propose workflows that separate representation learning from causal estimation to mitigate classical risks, such as the violation of identification assumptions. Further, new questions have been raised regarding challenges and validation requirements (e.g. Baumann et al. 2025; Bisbee and Spirling 2025; Hassan et al. 2025; Wood-Doughty et al. 2018). Much of this literature, however, remains anchored in disciplinary conventions and terminology unfamiliar to sociological audiences.

Against the background of this special issue's goal of revisiting foundational questions of explanation and causality in sociology in light of recent methodological developments, this article examines the intersection of text-as-data methods and causal inference. We first discuss current approaches to causal inference using textual

data by distinguishing between three central designs: when text is the outcome to be explained (text as outcome), when text functions as the treatment whose effects are to be estimated (text as treatment), and when information extracted from text is used as control variable (text as control). We then illustrate the text-as-outcome and text-as-treatment approaches through two empirical applications.[1] Throughout, we emphasise the central role of measurement uncertainty, from the data-generating process to model specification. Our aim is to contribute to an emerging methodological tool kit for text-based causal inference that remains consistent with the core commitments of analytical sociology. While we are primarily concerned with identification and estimation in causal inference using text as data, we also speak to broader questions of sociological explanation. In particular, we believe text-based representations can shed light on how treatments operate by revealing systematic changes in meaning, framing, or reasoning that mediate observed effects. At the same time, we do not claim that text-based methods uncover mechanisms automatically or exhaustively. Rather, they provide structured empirical access to intermediate processes that can support explanatory claims when combined with theory, design, and substantive interpretation.

2 What To Do with Texts?

Textual data has become a prominent source of empirical material in the social sciences (see, e.g., Gentzkow et al. 2019; Grimmer et al. 2022). As with other forms of data, text can enter a causal analysis in different roles: as an outcome, a treatment, or a control variable, and in some designs, even simultaneously in multiple roles.

In some designs, text is the object of explanation—the outcome of interest. Such designs focus on understanding why certain textual patterns emerge, such as the prevalence of particular topics or the use of specific frames. In other designs, text functions as the input, i.e. the treatment or stimulus that influences beliefs, attitudes, or behaviours. Finally, aspects of textual data can also serve as proxies or control variables, helping to adjust for confounding. Figure 1 illustrates the two main paradigms (text as treatment and text as outcome). On the left, text functions as a treatment: A machine learning (ML) system (or "ML assistant") creates or modifies text/image variants that are given to participants, and their subsequent behaviour is the outcome. On the right, text is the outcome: The occurrence of an event (represented by moon phases, such as the dark moon indicating no event and the full moon indicating an event) prompts individuals to produce text. This text is then analysed using ML assistants (using ML methods) to assess how those events influence textual expression.

[1] We focus on text-as-treatment and text-as-outcome because these designs are where sociologically relevant constructs do the most theoretical work; they are the quantities of interest around which causal identification is organised. Text-as-control is instrumentally useful but does not foreground the interpretive and measurement challenges that are central to our argument. Covering all three designs with equal depth would have overextended this paper.

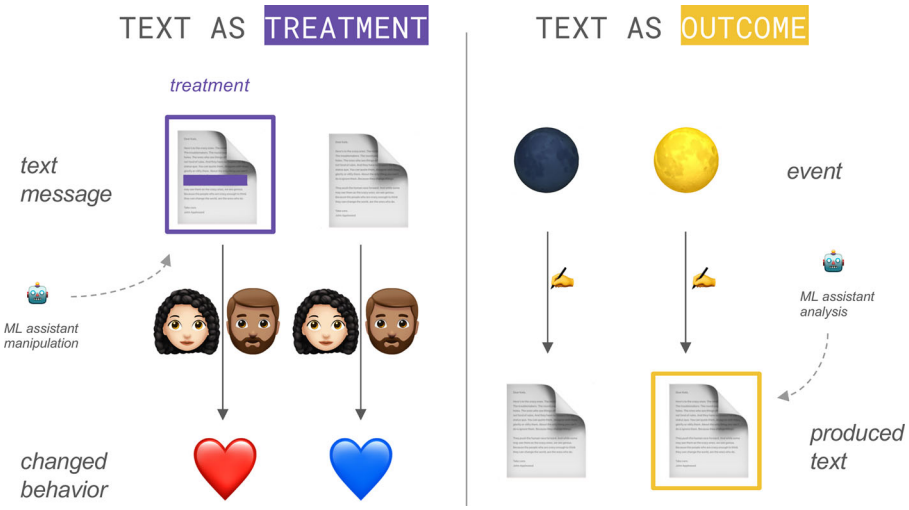

Fig. 1 Text as outcome and as treatment. *Note: ML* stands for machine learning.

In what follows, we focus on these main modes of incorporating text into causal analysis. In all cases, text cannot be analysed in its raw form. Rather, it must first be transformed, typically via a latent representation, into a format suitable for statistical inference (Egami et al. 2022; Feder et al. 2022; Grimmer and Stewart 2013; Jurafsky and Martin 2009).

2.1 Causality and the Potential Outcomes Framework

A central goal of sociology is to explain how and why social outcomes emerge and vary across social contexts, individuals, or time, linking assumptions about micro-level behaviour with macro-level conditions. Such explanations require both the identification of causal effects and the specification of the mechanisms through which these effects operate. Focusing on the identification of causal effects in this paper, we frame causal inference using the potential outcomes framework (Holland 1986; Rubin 1974, 2005; see also Jeon and Brand 2026, this issue). This framework formalises causality in counterfactual terms: The causal effect of a treatment is defined as the difference between the potential outcome under treatment and the potential outcome under control for the same unit. Formally, let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes for unit i, under treatment and control, respectively. The *causal effect* of a binary treatment $T_i \in \{0,1\}$ is defined as

$$\tau_i = Y_i(1) - Y_i(0)$$

This formulation highlights the fundamental problem of causal inference: It is impossible to observe both $Y_i(1)$ and $Y_i(0)$ for the same unit. One outcome necessarily remains counterfactual. Causal inference, therefore, depends on assumptions, such as unconfoundedness and the stable unit treatment value assumption (SUTVA), and on

research designs and statistical approaches (e.g. randomisation, matching, regression adjustment) that make these assumptions plausible.

Introducing text into the causal model complicates this structure. The potential outcomes framework presupposes that both the treatment $T_i \in \{0,1\}$[2] and the outcome Y_i are well-defined and measurable prior to analysis. In text-as-data applications, this assumption is often violated: Treatments and outcomes are not directly observed but must be inferred from unstructured text, requiring some reduction in complexity (Grimmer and Stewart 2013). Such reductions can take various forms, ranging from human coding and dictionary methods to scaling models and ML, and they generally occur as partly data driven. The boundary between discovery and estimation often blurs and creates new challenges, such as ill-defined causal estimands and overfitting. In the following sections, we discuss these issues in more detail and outline solutions, organised around the three scenarios of text as outcome, text as treatment, and text as control.

2.2 Text as Outcome

In the first setup, the outcome Y_i is derived from a text document such as a response to an open-ended survey question, an interview transcript, or a social media post. Unlike conventional numeric or categorical outcomes (e.g. income, voting behaviour, self-reported attitudes), outcomes are not directly observed but must be inferred from text. To map unstructured text into a structured, lower-dimensional representation, researchers rely on natural language processing techniques. These range from manual annotation by human coders to more automated methods such as dictionary-based classifications, probabilistic topic models, and vector representations from neural embeddings or large language models (LLMs; Grimmer and Stewart 2013; Imai and Nakamura 2024; Roberts et al. 2014). These methods reduce the complexity of text and extract patterns tractable for causal analysis. In sociological terms, they operationalise abstract constructs so that their empirical distribution can be modelled as an outcome variable. For example, one might estimate how a treatment (e.g. exposure to information about a social problem) influences expressed concerns or reasoning in texts. Formally, the mapping into a structured outcome space can be written as

$$Y_i = g\left(text_i\right) \text{ then estimate } Y_i \sim T_i + \boldsymbol{X_i}$$

Here, $g(\cdot)$ denotes the codebook function that maps unstructured text into labels or scores. The transformed Y_i is then explained by the treatment T_i and, where appropriate, covariates X_i. The act of dimensionality reduction—the definition of $g(\cdot)$—raises conceptual and methodological challenges (Egami et al. 2022). While theory can and should guide which categories or distinctions are substantively meaningful, it rarely provides a full specification of how linguistic features map onto them. In practice, the codebook function $g(\cdot)$ is usually developed inductively, through iterative refinement that combines human judgment with data-driven validation. This

[2] The treatment does not necessarily need to be binary.

makes it inherently corpus specific: Even when concepts are theoretically portable across contexts, their operationalisation via textual features is sensitive to domain, vocabulary, and structure.

A core methodological challenge arises because the same documents are used both to define the codebook function $g(\cdot)$ and to estimate causal effects (Egami et al. 2022): This dual use of the data heightens risks regarding identification and overfitting. Overfitting here refers not only to statistical artifacts from flexible modelling but also to the broader problem of researchers' degrees of freedom in defining $g(\cdot)$. Because text is highly expressive and unstructured, researchers face a vast space of potential representations: Different choices of tokenisations, embeddings, model architectures, labelling schemes, and thresholds can all shape the resulting variable. It becomes tempting to search over multiple candidate representations until one aligns with expectations or yields a significant result.

While not unique to textual analysis, the flexibility of text amplifies these risks. The boundaries between theory-driven coding and data-driven discovery are not clear-cut, and without a strict separation between the design of $g(\cdot)$ and the estimation of causal quantities, researchers risk inducing correlations that reflect overfitting rather than causal relationships. Moreover, such overfitting undermines the generalisability of findings. A codebook function that performs well on the data used to construct it may fail to capture stable patterns elsewhere.

A solution to mitigate overfitting and identification issues in learning the codebook function $g(\cdot)$ is to split the corpus into a training set and a test set for cross-validation (Egami et al. 2022). The training set is used exclusively to develop and validate $g(\cdot)$, whereas the test set is held out for the causal analysis. This separation ensures that any patterns discovered during training do not directly inform the estimation of causal effects, thereby preserving out-of-sample validity. By isolating codebook learning from causal inference, sample splitting improves robustness, enhances portability, and guards against spurious results (see also Breen and Pan 2026, this issue, on the relevance of external validity).

2.3 Text as Treatment

In alternative research designs, text can serve as the treatment. Here, the treatment is often a latent feature of the text. Using text in this way is already common in the social sciences. Vignette experiments typically rely on textual descriptions of hypothetical scenarios as treatments (Treischl and Wolbring 2022).

Formally, let $T_i = g(\text{text}_i)$, where $g(\cdot)$ is a model extracting the treatment from text. Causal inference proceeds by estimating the effect of T_i on the outcome Y_i as follows

$$Y_i = \beta T_i + \epsilon_i$$

Although vignette experiments are widely used to study causal effects in complex social scenarios, researchers rarely acknowledge that treatments are latent and manipulable only through changes to the wording of the text. It is assumed that $g(\cdot)$ is fully known. However, the fact that treatments are latent creates a methodological

challenge: Manipulations intended to isolate a specific treatment dimension may inadvertently vary other unmeasured features of the text, thus violating assumptions necessary for causal identification. In particular, causal identification relies on the no-aliasing assumption, assuming that the effect of any unmeasured latent treatment that covaries with the measured treatment is zero, and the no-interaction assumption, assuming that the effect of the measured treatment is independent of unmeasured treatments. Both of these assumptions are at risk in text-based designs, which raises concerns about both internal and external validity. One strategy to address these concerns is to design experiments with many vignettes per latent treatment, which allows testing of assumptions and statistical adjustment for latent confounders. This approach aims to disentangle the target treatment effect from spurious variations in the text.

To facilitate this adjustment, recent work has introduced probabilistic models that infer latent features from text, such as the supervised Indian buffet process (SIBP; Fong and Grimmer 2016), a non-parametric Bayesian method that represents each document as a combination of multiple latent features. By jointly modelling text, treatment, and outcome, SIBP allows multiple correlated latent dimensions of the text to be inferred simultaneously, rather than conditioning on a single, researcher-defined treatment (e.g. wording changes in vignettes). This reduces violations of the no-aliasing assumption by explicitly modelling co-occurring latent treatments and mitigates no-interaction concerns by allowing outcomes to depend on combinations of latent features rather than on an isolated treatment dimension. Treatments do not need to be predefined; instead, candidate latent treatments can be discovered directly from text. A more detailed discussion of the SIBP can be found in Sect. A.1 of the appendix.

2.4 Text as Control

In a third class of research designs, text can work as a source of information for adjusting for confounding (Roberts et al. 2020; Imai and Nakamura 2024, 2025). In many observational settings, treatment assignment and outcomes are jointly influenced by latent characteristics that are not directly observed but are expressed in language. Textual data can therefore act as a rich proxy for otherwise unmeasured confounders. Formally, let Z_i denote a set of latent confounders affecting both treatment, T_i, and outcome, Y_i. When these confounders are partially encoded in text, researchers can construct control variables as

$$Z_i = g\left(\text{text}_i\right)$$

and estimate causal effects under the assumption of conditional ignorability,

$$Y_i\left(t\right) \perp T_i \mid g\left(\text{text}_i\right), X_i$$

with $g\left(\cdot\right)$ mapping unstructured text into a lower-dimensional representation that captures confounding variation relevant for both treatment assignment and outcomes.

Using text as a control variable is appealing because language often reveals information that is difficult to measure through structured covariates alone. For example, political ideology, emotional states, policy priorities, or socioeconomic background may be expressed implicitly through word choice, framing, and argumentation. In such cases, textual representations can reduce omitted-variable bias by absorbing variation correlated with both treatment and outcome. Recent advances in representation learning have expanded the tool kit for using text as a source of confounding control. Unsupervised and supervised topic models, document embeddings, and generative models can be used to learn latent structures that summarise confounding information at scale (Roberts et al. 2020; Imai and Nakamura 2024, 2025).

Although these developments are promising, text as control poses the methodological challenges discussed above due to its dimensionality and flexibility. In the following empirical examples, we will focus on text as treatment and text as outcome.

3 Empirical Applications

3.1 Text as Outcome

The first application uses open-ended survey responses as the outcome of interest. The goal of the analysis is to estimate the causal effect of survey stimulus dimensions on latent features of the textual response.

3.1.1 Introduction and Theoretical Background

The presented data in this study come from a research project conducted by the first author, designed to study the activation of criminality-related stereotypes in visual perception in fast- and slow-thinking settings (Kahneman 2011; Tutic et al. 2024). Theoretically, the study builds on the ideas that individuals hold pre-existing notions of who commits crimes and what a "typical" criminal looks like, and that decisions are shaped not only by the factual content of a scene but also by underlying social stereotypes, particularly in ambiguous situations (Bull and Green 1980; Correll et al. 2002; Eberhardt et al. 2004; MacLin and Herrera 2006; Payne 2001). Following previous empirical results, which have shown that Black (see, e.g., Duncan 1976; Eberhardt et al. 2004) and Arab/Muslim men (see, e.g., Stelter et al. 2023) are particularly often stereotyped and considered more criminal or violent, we expected individuals to label these sociodemographic groups more often as criminal in ambiguous depictions.

We conducted a factorial survey experiment with visual vignettes in which respondents described images of ambiguous crime scenarios in open-text answer fields. The crime scenarios showed offenders of different ages and races/ethnicities. We assumed that in this context, cues regarding age and ethnicity/race may affect whether a behaviour is interpreted as suspicious or deviant. In our analysis, we distinguish between descriptive responses, which neutrally describe the scene or the actor without evaluative judgment, and criminalising-normative responses, which express judg-

ments about the potential criminality. By mapping these latent interpretive categories from text, we can assess how the experimental manipulations affect stereotype activation in participant narratives.

Variation in normativity across conditions may signal bias and prejudice in the interpretation of identical scenarios. Our research question is thus focused on understanding how visual characteristics of the depicted person (in terms of age and race/ethnicity) affect whether the respondent describes the scene in criminalising-normative versus descriptive terms.

3.1.2 Experimental Design

The study employs a factorial survey experiment with visual vignettes, showing participants photorealistic, artificial intelligence (AI)–generated images of ambiguous crime scenarios. Vignettes allow systematic manipulation of theoretically relevant features—here, the presumed offender's age and ethnicity/race—while holding other aspects of the scenario constant.[3]

Half of the participants were placed in a reflective decision-making condition in which they first described the image in a few sentences before making further judgments on closed-ended scales, whereas the other half responded spontaneously without in-depth engagement. Each participant viewed three pre-tested images, one from each of three different ambiguous scenarios (Fig. 2). In this study, we focus on the descriptions provided by participants in the reflective condition.

Each participant viewed one image per scenario. For each scenario, the participant was randomly assigned to one of multiple image versions in which the offender varied systematically in age (young/older) and phenotype (Black, Arab, White), with presentation order fully randomised across participants.

3.1.3 Data and Codebook Function

This application uses data from a factorial survey experiment embedded in the 76th wave of the German Internet Panel, a longitudinal study based on a random probability sample of the general population in Germany aged 16 to 75 years (Blom et al. 2015); open-text answers are accessible only through the on-site data access. The survey included 3423 respondents, half of whom ($n = 1719$) were encouraged to provide reflective descriptions, yielding $n = 1719 * 3 = 5157$ image descriptions. These responses average 18.58 words ($median = 14$, $min = 0$, $max = 231$). The outcome of interest is whether a text is descriptive or criminalising-normative. Descriptive texts report what is visible or happening without overt judgment or normative inference. Criminalising-normative texts include evaluations, suspicions, moral judgments, and implications of (in)appropriateness. Our goal is to estimate the causal effect of randomly assigned image features on the likelihood of a criminalising-normative interpretation.

[3] The focus of the research project was on ethnicity/phenotype; please note that the older and younger presumed offenders vary in more dimensions, such as social status.

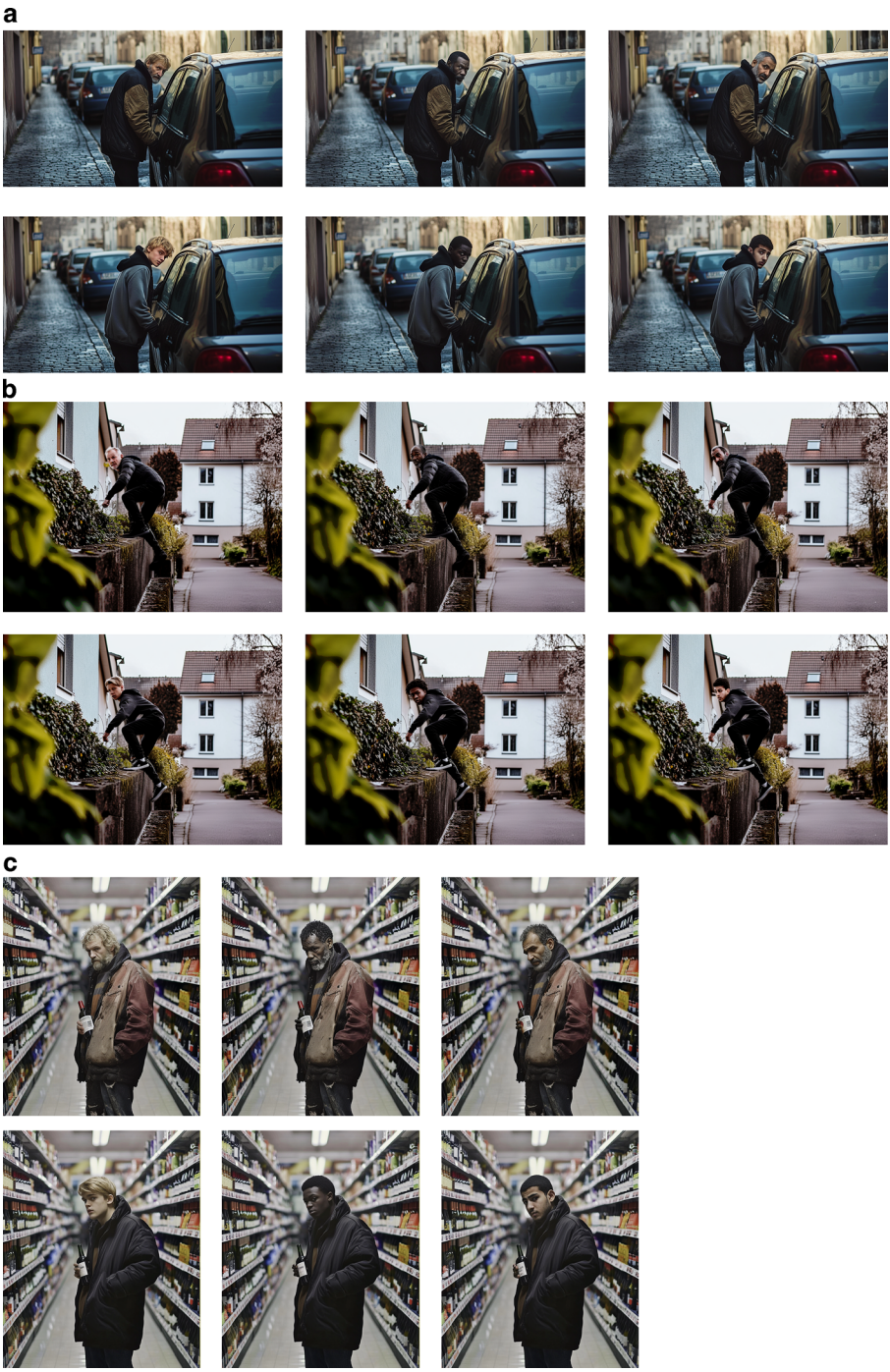

Fig. 2 Stimulus images from the image description task. **a** Scenario 1: car. **b** Scenario 2: house. **c** Scenario 3: shopping

To derive a structured, low-dimensional outcome from unstructured text, we implemented a manual coding approach, classifying responses into binary categories (*criminalising-normative* versus *descriptive*) based on semantic content.

Step 1: Sample Splitting. Following Egami et al. (2022), we split the corpus ($n=$ 5078 non-missing responses) into a training set and a test set to separate outcome definition from causal analysis. The training set is used exclusively to develop and validate the coding scheme, and the test set is held out for final causal estimation, ensuring out-of-sample evaluation and avoiding overfitting.

Given that outcome coding was fully manual in our case, we allocated 10% of the data to the training set and 90% to the test set. This maximises the number of observations for causal estimation. Larger training sets could allow for more complex codebooks and might be particularly important when employing unsupervised or weakly supervised strategies.

Step 2: Codebook Development. Within the training set ($n=508$), a subsample of $n=307$ responses was used to iteratively develop a coding protocol following a qualitative content analysis approach that combines theory-driven category construction with inductive refinement. This translates the theoretical concept of "normative judgment" into specific labelling rules (e.g. Mayring 2015). Coding criteria were theory informed (i.e. by clarifying what normative judgment entails) but refined inductively through close reading and discussion of example responses. This involved identifying semantic cues, edge cases, and ambiguous formulations to refine the boundary between criminalising-normative and descriptive categories with the empirical material. This step constitutes the substantive foundation for the codebook function g(·): a mapping from text to a theoretically meaningful, low-dimensional outcome variable. This approach highlights how classical content analysis of open-ended responses can be applied within the context of causal inference. In the course of coding, we identified a recurring pattern: Many responses did not fall neatly into a single category but instead referenced multiple possible interpretations of the scene. To systematically capture this nuance, we introduced an ambiguity flag marking responses with competing or layered interpretations. Each response was assigned one of two primary labels and potentially flagged for ambiguity:

- *Criminalising-normative (1):* Expresses judgment, value, or social/moral norm. Includes evaluative language, suspicions, assumptions about intentions (e.g. theft), and comments on appropriate behaviour or calls to action. Makes reference to criminal behaviour.
- *Descriptive (0):* Neutral observations or factual descriptions without a clear evaluative preference. May include interpretive distance or meta-commentary (e.g. "The image seems to suggest …").
- *Ambiguity flag (1/0):* Indicates whether multiple interpretations or alternative readings are present. General uncertainty (e.g. "maybe", "perhaps") alone does not trigger the flag.

A conservative approach was applied: A response was coded as criminalising-normative only when that frame was clearly dominant; ambiguous responses were coded as descriptive but flagged as ambiguous.

Step 3: Codebook Refinement. The initial coding scheme was developed by the primary author and refined through annotation by trained student assistants with prior experience in data annotation (two 25–30-year-old master's degree students in sociology; they are both native German speakers without a migration background). Initially, assistants worked on the training set used to develop the codebook to align interpretations, discuss ambiguous items, and clarify coding instructions.

The remaining training set ($n = 201$) was independently coded by all three coders to validate consistency and refine guidelines. Krippendorff's alpha (nominal) was computed across three dimensions to evaluate interannotator agreement (IAA):

- Normative label: $\alpha = 0.742$
- Ambiguity flag: $\alpha = 0.671$
- Normative-mentioned indicator (i.e. responses flagged as either normative or ambiguous): $\alpha = 0.866$

We used the normative-mentioned indicator as the outcome for downstream analysis. This broader category ensures higher intercoder agreement while remaining conceptually aligned with the theoretical goal of identifying normatively charged interpretations.

Step 4: Applying the Codebook via Human Annotation. Once the coding scheme was fully developed in steps 2 and 3 and thus arriving at the coding rules presented in step 2, the assistants annotated the full dataset, i.e. the test set, according to the finalised codebook. This produced a structured outcome variable derived entirely from held-out data. This human-in-the-loop approach ensures high fidelity to the theoretical constructs, as coders apply a refined and context-sensitive codebook developed through iterative discussion. Automated methods, such as classifiers, can scale to larger corpora but may underperform when concepts are contested, are context dependent, or require interpretability. Human coding remains a gold standard, albeit with limited scalability (however, see Bisbee and Spirling 2025 for current discussions). Next, we proceed to causal modelling.

3.1.4 Causal Modelling

Following the procedure described above, we derived the outcome $\widehat{Y}_i$. Causal effects were then estimated using standard logistic regression with cluster-robust standard errors:

$$\widehat{Y}_i \sim T_i,$$

where T_i represents the treatments (age, ethnicity, scene) and Y_i represents the normative-mentioned binary indicator. The results of the logistic regression are visualised in Fig. 3.

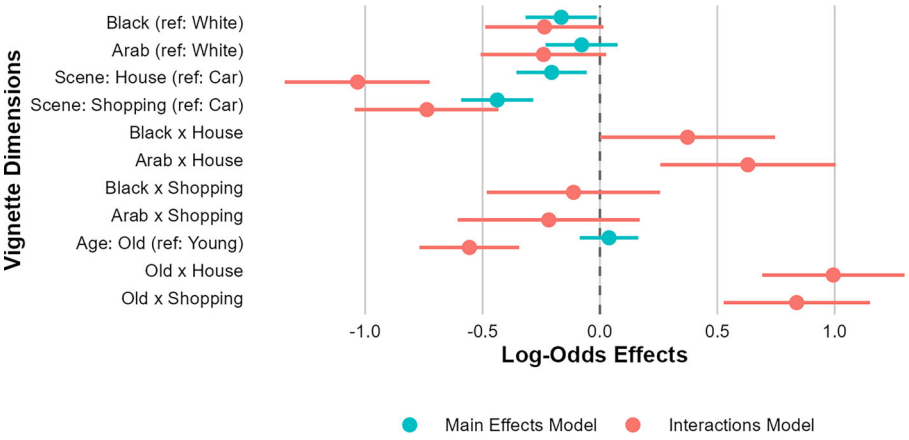

Fig. 3 Coefficient plot for the models on normative mention. Logistic regression with clustered standard errors. Data: German Internet Panel, wave 76

Overall, the models reveal scene-dependent patterns in how sociodemographic markers of the depicted offender influence the likelihood of a normative interpretation. In the main effects model in Fig. 3, which includes only main effects, offenders depicted as Black or Arab were somewhat less likely than White offenders to elicit normative mentions, with a statistically significant effect for Black individuals. The car scene was most likely to result in criminalising-normative interpretations, and there is no significant difference by age of offender.

To explore heterogeneity by scene, we added interaction terms in the interactions model. The results show that the ethnic penalty varies across settings. In the car scene, White individuals were significantly more likely to be described in normative terms, whereas this racial/ethnic penalty diminishes or even reverses in the other contexts: For instance, Arab men were more likely than White men to elicit normative interpretations related to burglary or trespassing in the house scene. Regarding age, older offenders were significantly more likely to be described in criminalising-normative terms in the house and shopping scenes (compared to the younger offender) but were significantly more likely to be described purely descriptively in the car scene.

3.1.5 Discussion

In this empirical application, survey respondents were presented with images of ambiguous crime scenarios and asked to describe them. The depicted offenders varied systematically by age and ethnicity/race. Our results align with theoretical expectations that social perceptions of deviance and criminality are shaped by the interaction of demographic cues and situational ambiguity and may also be influenced by social desirability biases. We also observe context-dependent effects: The shopping scene elicited fewer normative responses overall, and the car and house scenes activated different implicit stereotypes. These patterns are consistent with responses to closed-ended survey questions collected in the same study, providing

robustness to this text-as-data analysis, and vice versa (see descriptive results in Sect. A.2 in the supplementary material).

This application illustrates that normative language in textual responses can serve as a meaningful signal of latent social cognition in response to controlled visual stimuli. Key features of our approach are the human-in-the-loop coding procedure and the train/test split: Trained student assistants applied a codebook to classify open-ended responses. The codebook was developed using only a subset of the data, which was not used in downstream causal analysis. Without this separation, there would have been a danger that dimensions of the stimulus itself would become inadvertently baked into the codebook as cues for normative language. If those same responses were then used again for causal estimation, we would risk conflating properties of the stimulus with the expression of social norms, inflating apparent effects.

Importantly, although our analysis treats textual responses as outcomes, the stimuli themselves—AI-generated images—also embed latent constructs. That is, respondents interpret not just explicit features such as ethnicity/race or age but also subtle visual cues embedded in the image-generation process. This latent nature of visual stimuli introduces complexity and highlights the need for future research to better model and decompose image features (and/or apply stimulus sampling), an aspect ignored in this application (see also Fong and Grimmer 2023; Schwitter 2025).

3.2 Text as Treatment

The second application examines a design in which the treatments are unknown latent features embedded in texts. The empirical goal is to discover these unknown latent treatment candidates from the texts and simultaneously estimate their effects on an outcome.

3.2.1 Introduction and Theoretical Background

Our application investigates whether exposure to first-person accounts of mothers' experiences with household and childcare responsibilities influences fathers' fairness perceptions of the division of such responsibilities. The intervention builds on research on perspective-taking and perspective-getting, which involve imagining another person's point of view or learning directly about it. These approaches have been shown to reduce prejudice and foster empathy in contexts such as ethnic discrimination and political polarisation (Broockman and Kalla 2016; Galinsky and Moskowitz 2000; Kalla and Broockman 2023). Whereas prior work has largely focused on large social distances between groups, this study applies perspective-taking to a low-distance, high-role-similarity context: Mothers and fathers of young children share a common role yet often experience unequal distributions of household tasks (see, e.g., Haupt and Gelbgiser 2024). Inequality in this setting tends to appear in subtle and normalised forms embedded in everyday practices and normative expectations, resulting in comparably higher household responsibilities,

cognitive household labour, and mental load for mothers (Daminger 2019; Haupt and Gelbgiser 2024; Hochschild and Machung 1989).

From the perspective of social role theory (Eagly 2013), gendered divisions of labour produce and reinforce expectations that men act as breadwinners and women as primary caregivers. The role prioritisation model (Haines and Stroessner 2019) predicts that norm violations, such as men engaging intensively in childcare, can trigger backlash. Reframing such behaviour in ways that affirm a shared commitment to both work and family may reduce these reactions. In this context, perspective-taking narratives may shift the salience from gender categories to the shared identity of "parent", thereby reducing resistance to more equal divisions of labour. We expect that exposure to first-person accounts of mothers' experiences with household and childcare responsibilities will make the often invisible aspects of these responsibilities more salient to fathers, resulting in changes in their (fairness) perceptions of household and childcare division. Following the empirical goal described above, we aim to identify which aspects might be specifically relevant in mothers' narratives to cause changes in fathers' perceptions and beliefs.

3.2.2 Experimental Design

We designed a survey experiment in which 984 fathers of young children were exposed to a randomly selected text stimulus, mimicking a social media post of a mother on a parenting subforum on Reddit, a popular social media platform (see Amaya et al. 2021 for details about the Reddit platform). Participants were recruited via the online platform Prolific and screened according to specific eligibility criteria: They had to be fathers of at least one child born between 2015 and 2025, be living with their children and partner, identify as heterosexual, have English as their primary language, and be located in the United Kingdom or the United States. These criteria ensured that perspective-taking operated in a low-distance, high-role-similarity context (fathers reading about mothers' experiences).

A total of 200 textual stimuli were formatted to resemble a mother's post on the *r/Mommit* subreddit (see Fig. 4 for an example); r/Mommit is a subforum on Reddit dedicated to fostering exchange and discussion among mothers. To avoid ethical and practical issues with using third-party social media content, we created the 200 posts using a controlled text-generation procedure with the LLM GPT-4.1. GPT-4.1 was prompted to produce realistic Reddit posts of approximately seven sentences in length, written in the first person and accompanied by three short supportive comments. The supplementary material provides the full prompt used (Sect. A.3).

The generation procedure aimed to produce variation along multiple dimensions, including emotional intensity, focus on individual versus systemic issues, partner involvement, and thematic content such as meal planning, bedtime routines, childcare costs, and developmental milestones. Moreover, we instructed the LLM to vary the content as either emphasising challenges and invisible burdens or highlighting joys and positive parenting moments, and the style was either a personal narrative that presented a clear perspective or a factual description that was less personalised. The texts also varied along multiple uncontrolled stylistic and thematic dimensions, such as emotional valence, the extent of partner comparison, the presence of moral

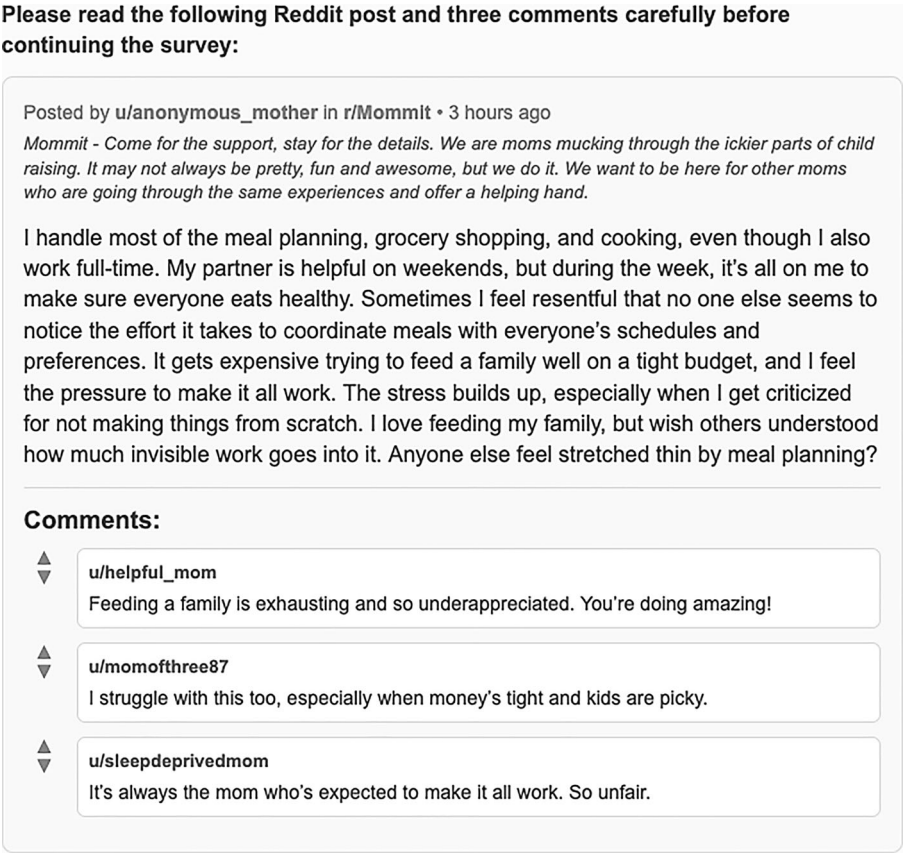

Fig. 4 Example of a (simulated) post to r/Mommit

framing, the relative emphasis on burden versus joy, and references to systemic constraints. All dimensions are not directly observed, i.e. latent, and must therefore be inferred from the data using, e.g., the SIBP procedure (Fong and Grimmer 2016). In our example, all treatments are unknown and are learned from the texts; while this is one approach to treatment identification, in many settings researchers might have theoretical expectations about relevant treatment dimensions, and hybrid approaches, in which some dimensions are specified ex ante and others are discovered from the data, are also possible.

3.2.3 Outcomes

Following exposure to the randomly assigned post, respondents completed two short multi-item measures capturing (i) respondents' own normative beliefs towards mothers' responsibilities in household and childcare tasks, and (ii) perceived fairness of the division of household and childcare responsibilities in their own family. Each construct was built as an additive index of four items. We measure respondents'

Table 1 Items for personal beliefs (*PB*) and perceived fairness (*PF*) scales

Scale	Item sescription
PB	Mothers should do most of the household work (cleaning, cooking, laundry, grocery shopping)
	Mothers should do most of the childcare work (feeding, bathing, homework, bedtime)
	Mothers should take main responsibility for planning/organising routine household chores (keeping track of supplies, planning meals)
	Mothers should take main responsibility for planning/organising childcare (doctor appointments, school events, extracurriculars)
PF	How fair is the division of household work between you and your partner?
	How fair is the division of childcare work between you and your partner?
	How fair is the division of planning/organising household chores?
	How fair is the division of planning/organising childcare?

views on mothers' roles and household division of labour using two scales: *personal beliefs (PB*; a 5-point scale ranging from 1 = strongly agree to 5 = strongly disagree) and *perceived fairness (PF*; a 5-point scale ranging from 1 = very unfair to me to 5 = very unfair to my partner). Table 1 summarises the items for each scale. We expect effects on both outcomes as perspective-taking should make mothers' (often) comparatively greater responsibilities, including the otherwise invisible aspects of cognitive household labour and childcare, more visible to fathers.

3.2.4 Learning Latent Text Treatments with SIBP

To identify the latent features of the textual posts and measure their impact on the survey outcomes, we use the SIBP by Fong and Grimmer (2016; see the appendix for a detailed description). The SIBP models each text as a combination of multiple unobserved binary features, such as explicit partner comparison, expression of emotional exhaustion, or emphasis on systemic barriers. The model incorporates the observed outcomes to prioritise features that are predictive of variation in the outcomes. The resulting binary feature–document matrix, estimated using only the training data, is then used in a regression framework to estimate the causal effects of these features on the two outcome indices (using the test data). Causal effects are quantified as average marginal component effects (AMCEs; Hainmueller et al. 2014). Intuitively, the AMCE for a given feature represents the average difference in the outcome when that feature is present in a text compared to when it is absent, averaged over the distribution of all other features in the corpus.

The main challenge with the SIBP approach is selecting the model that uncovers the most substantively meaningful latent treatments. To address this, researchers should run a series of models with varying parameter configurations (Fong 2019) and evaluate them using the coherence metric proposed by Fong and Grimmer (2016, p. 1605), which helps identify parameter configurations likely to reveal the most interesting treatments to investigate further. The choice of the final parameter configuration should be made based on which latent textual treatments are substantively the most interesting and relevant (Fong 2019).

Table 2 Latent treatment interpretations with representative top words for personal beliefs (*PB*) and perceived fairness (*PF*) models

Model	Latent treatment variable	Top words
PB	Milestones and learning	*Bike, training, wheels, cheering, rode, witness*
	Event planning and birthdays	*Details, birthday, pull, cards, times, it's*
	Daily routines and moods	*Afternoons, weekday, contagious, cracked, makes, bread*
	Meal planning and food work	*Planning, eats, hours, preferences, cooking, grocery*
	Play and creativity	*Perfect, met, allergies, cardboard, mini, food*
	Household management and mental load	*Checklist, formula, fridge, hint, hoping, inventory*
	Motherhood identity and emotions	*Motherhood, clashing, sitting, sounds, laughing, mom*
	Childcare logistics and career	*Backup, flexible, waitlists, childcare, career, nanny*
PF	Reflections and narratives	*Stories, expected, chao, incredibly, watching, daughter*
	Expectations and struggles	*Wrong, expect, invisible, keeping, family, worrying*
	Childcare coverage and stress	*Stress, backup, coverage, negotiable, childcare, costs*
	Daycare and finances	*Budget, daycare, contagious, insisted, call, celebrate*
	Infant care and basics	*He'll, request, diapers, communication, job, baby*
	Partners and support networks	*Excuse, offer, splitting, he'd, extended, gatherings*
	Milestones and achievements	*Bike, training, wheels, cheering, contagious, wobbled*
	Planning and routines	*Prep, thoughtful, zoom, family's, couldn't, blur*

For our study, we estimated a series of SIBP models separately for each outcome (PB and PF). In each scenario, we manually inspected the words most strongly associated with the latent treatments to select the models for downstream analysis. In both models, we set the number of latent treatments to be discovered to eight[4] and split the data 50/50; that is, half of the data was used for latent treatment discovery and half of the data was used for AMCE estimation. Following common practice for bag-of-words approaches, we removed stopwords from the texts and trimmed rare words (in our case, words that appeared less than seven times in the corpus).

3.2.5 Results

Per our specification, the SIBP uncovers eight latent treatment dimensions for both outcomes (PB and PF). Table 2 presents the treatment labels that we developed by qualitatively interpreting the top words associated with each latent feature in light of the theoretical background. The recovered topics capture a wide range of everyday parenting experiences and associated forms of cognitive household labour, as we would expect given the prompts used to generate them with the LLM.

[4] Setting the number of treatments to eight is a somewhat arbitrary choice, but this number led to a good balance between model complexity and interpretability.

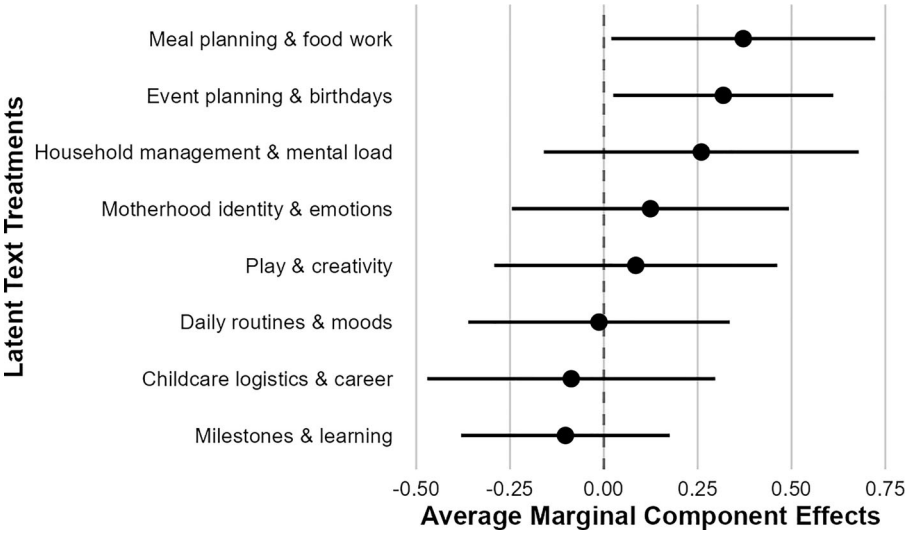

Fig. 5 Coefficient plot for the model on personal beliefs

For PB, treatments emphasise concrete parenting practices and emotional investments; for PF, the latent treatments focus more on normative and evaluative dimensions of parenting.[5] Taken together, the results show that our models successfully recovered meaningful latent treatment dimensions from the first-person accounts of mothers' experiences with household and childcare responsibilities. Learned latent treatments are both substantively meaningful and theoretically aligned with the literature on cognitive household labour and gendered division in household and childcare labour. They highlight how everyday practices such as meal planning or bedtime routines coexist with broader themes of identity, fairness, and invisible labour.

In the next step, we link these features to the outcomes (PB/PF) to determine their causal impacts on fathers' perceptions and beliefs. Figures 5 and 6 display the estimated AMCEs for each outcome. Overall, we find modest evidence that perspective-taking through exposure to first-person accounts of mothers' experiences with household and childcare responsibilities influences fathers' perceptions and beliefs. For PB, we find that the latent treatments "meal planning and food work" as well as "event planning and birthdays" lead to stronger disagreement with the statements that mothers should be primarily responsible for household and childcare tasks. For PF, exposure to the "milestones and achievements" treatment decreases fathers' perceptions that household and childcare responsibilities in their family are unfair towards their partners. That is, fathers tend to perceive that household and childcare responsibilities in their family are more unfair towards themselves after

[5] Note that the text corpus used for learning the latent treatments is the same for both models (PB/PF). However, due to SIBP's focus on discovering features that are predictive of the outcome, treatments may vary among different outcomes, even when using the same input data.

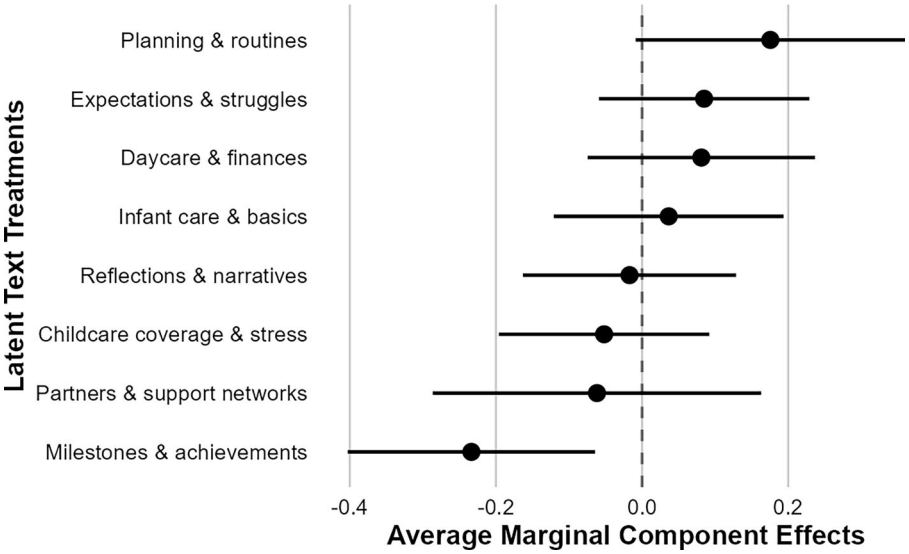

Fig. 6 Coefficient plot for the model on perceived fairness

being exposed to this treatment. The majority of latent text treatments do not seem to cause changes in PB or PF evaluations, however.

3.2.6 Discussion

The results of this application demonstrate that latent treatment discovery from textual data with SIBP can successfully recover dimensions that resonate with established sociological descriptions of cognitive household labour and the gendered imbalance in household responsibilities and childcare. Treatments such as meal planning, birthday organisation, and bedtime routines closely mirror the "invisible" forms of planning and coordination identified by previous work (Daminger 2019; Haupt and Gelbgiser 2024). This is a non-trivial achievement: Even though the texts were generated in a controlled manner to vary along these dimensions, the fact that the SIBP can recover them shows that probabilistic latent feature models can meaningfully detect theoretically relevant constructs from short, conversational posts. In this sense, the application highlights the methodological feasibility of using text-as-treatment designs to capture subtle and often hidden aspects of household labour.

At the same time, the causal effects of these treatments on fathers' perceptions are modest. The perspective-taking hypothesis that brief exposure to mothers' voices makes invisible responsibilities more salient and leads to more egalitarian fairness judgments receives only limited support in this design. This may be due to deeply entrenched gender norms, especially around the cognitive and invisible dimensions of labour, which may explain why perceptions and beliefs shift only a little through one-time brief exposures in an experimental setting. Overall, our study should be read less as a definitive test of the effectiveness of perspective-taking interventions

and more as a proof of concept for the methodological strategy of discovering latent treatments from text data.

From a methodological standpoint, the application highlights the interpretive leeway inherent in SIBP. Although the framework cleanly separates the discovery of latent treatments from the estimation of their effects, substantial researchers' degrees of freedom remain in selecting model configurations, labelling treatments, and deciding which features are substantively interesting. The statistical formalism of the method can obscure the extent of this qualitative judgment, making transparency about these choices essential. In this sense, the promise of the approach lies as much in its ability to structure interpretive work as in its statistical output.

Looking ahead, future methodological work could explore the integration of generative AI–powered causal representation learning approaches, such as those proposed in Imai and Nakamura (2024, 2025), to allow latent treatment discovery and effect estimation with state-of-the-art representation learning architectures and joint modelling frameworks. Such developments hold promise for reducing the gap between highly controlled experimental stimuli and the richness of naturalistic texts, while maintaining rigorous identification strategies.

4 Scalability of Open-Answer Coding

In our first use case, the factorial survey experiment, we have more than 5000 open-text image descriptions (and in other studies, researchers might have many more). Even after refining the codebook and validating IAA on a training subset, scaling manual annotation to the full dataset would require substantial coder time and repeated adjudication (Von der Heyde et al. 2025). This problem is typical for open-ended survey responses and many other social science open-text answers: Annotation cost grows with sample size, and causal inference studies often benefit from large samples in terms of statistical power and improved subgroup analyses. Increasingly, (L)LMs provide a scalable solution by acting as automated annotators that can produce consistent labels at negligible marginal cost.[6] Recent work shows that LLMs can achieve annotation quality comparable to that of crowd workers and, in some cases, trained annotators when tasks are well-specified and supported by a clear codebook (Gilardi et al. 2023; Schonlau et al. 2023; Törnberg 2023; Halterman and Keith 2025).

We initially tested the comparability of an LLM-based survey response labelling method against the gold standard coded by two student coders. Following Halterman and Keith (2025), we kept the prompt identical to the codebook for consistency—although shorter prompts are more efficient, we prioritised interpretability and comparability. Our prompt for coding survey responses mirrors the codebook used in our first case (Sects. A.4 and A.6 in the supplementary material). Using

[6] However, the growing adoption of machine learning models also raises concerns about computational costs and its consequences, such as increased carbon footprint, especially as these models become larger and demand more resources and impact (Strubell et al. 2019; Luccioni et al. 2024a, b, Fernandez et al. 2025).

GPT-5.2 (February 2026) from OpenAI as a state-of-the-art model (LLM label), we benchmarked performance by comparing the LLM-generated label as a new coding instance against human labels and measuring agreement. The LLM labels (with $\text{std}_{\pm}$ based on three re-runs) show good concordance with human coding (0-shot IAA kappa = 0.63 $_{\pm 0.01}$ or 3-shot kappa = 0.78 $_{\pm 0.01}$ for annotating the *normative flag*),[7] indicating some open challenges in understanding the coding scheme (Sect. A.5 in the supplementary material). Von der Heyde et al. (2025) also discuss some limitations of this approach and the advantages of an adapted fine-tuning modelling approach. To reliably validate and benchmark our use case, we propose establishing a *standard test set* as the gold standard for validating and calibrating our survey codebook. To balance scalability and validity, we use a hybrid pipeline:

- *Human gold standard:* A subset of responses is independently coded by (trained) annotators.
- *LLM synthetic labelling:* The LLM labels the full dataset using the established codebook (e.g. Halterman and Keith 2025).
- *Evaluation:* Synthetic labels are compared to gold labels using agreement metrics.
- (Optional) *Selective review:* Disagreements or ambiguous cases are manually adjudicated, drawing on disagreement-aware annotation principles (e.g. Uma et al. 2021; Röttger et al. 2022; Weber-Genzel et al. 2024; Xu et al. 2025).

This method significantly lowers annotation costs while maintaining measurement accuracy. Nonetheless, the different validation phases carried out by humans ensure human involvement in the coding process and provide a consistent reference point for control. Additionally, it should ensure a replicable LLM pipeline (Baumann et al. 2025; Barrie et al. 2025).

Distinguishing Annotation Scalability from Synthetic Survey Responses. Scaling our second use case requires additional survey responses. There is a growing literature using LLMs as survey respondents or generators of synthetic opinion data. For example, studies simulate public opinion or social behaviour by prompting LLMs with demographic *personas* (Argyle et al. 2023; Aher et al. 2023; Hu and Collier 2024; Kaiser et al. 2025). While these approaches are useful for exploring hypothetical populations, they do not solve the scalability problem addressed here. Our goal is to code real survey responses to recover causal effects in observed populations, not to simulate new respondents. Using LLMs to generate synthetic survey responses requires models to capture heterogeneous perspectives across demographic groups. Achieving this reliably would require validated persona simulations and calibration against real survey distributions, which remains an open research problem (Liu et al. 2024; Beck et al. 2024; Sun et al. 2025). Existing work shows that LLM-generated opinions often reflect training-data biases or lack subgroup realism (Santurkar et al. 2023; Huang et al. 2024; Xu et al. 2025). Xu et al. (2025) show that LLM-generated annotations using zero-shot persona prompting only moderately align with human ratings. This alignment varies significantly across different tasks and demographic

[7] Cohen's kappa is used here because it measures agreement between two annotators. Krippendorff's alpha is more appropriate for situations involving multiple annotators.

groups. Their research also indicates that blindly adding synthetic annotations during model training can be detrimental, with the best blending strategies heavily dependent on the nature of disagreements within the data. Consequently, although synthetic responses may help with hypothesis testing or data augmentation in settings with high disagreement, they are not yet a reliable replacement for scalable annotation of actual survey data.[8]

5 Measurement Uncertainty

As our previous examples have also shown, text-as-data approaches introduce multiple layers of measurement uncertainty that compound throughout the causal inference pipeline. These challenges stem from the complexity of natural language (generation), the opacity of computational tools, and the interpretive judgments required to map unstructured text onto theoretical constructs (Wood-Doughty et al. 2018; Zhang and Zhang 2022; Baumann et al. 2025; Bisbee and Spirling 2025; Hassan et al. 2025; Modarressi et al. 2025). In the following, we discuss measurement issues from our two empirical applications presented previously.

Synthetic and Natural Text-as-Data Collection. When researchers use natural or synthetic (generated) content to create experimental stimuli, whether textual posts or visual materials, they introduce dependencies on quality, consistency, and transparency. In our applications, both synthetic images (depicting ambiguous crime scenarios) and synthetic Reddit posts created measurement challenges. State-of-the-art models and multimodal image generators are often closed-source (e.g. OpenAI-API, Midjourney), limiting reproducibility and researchers' ability to verify that generated materials adequately represent intended theoretical constructs. This implies that the management and reproducibility of a single generation process, such as prompting constraints, restrict scientific robustness (Barrie et al. 2025). For the text-as-outcome study, synthetic images showing individuals of different ages and races/ethnicities may contain subtle, unintended variations beyond the manipulated dimensions. These latent visual features could influence respondents' descriptions in ways that confound the intended treatment effects. Similarly, synthetic Reddit posts may contain systematic linguistic patterns or artefacts from the generation process that differ from authentic user-generated content, potentially limiting external validity.

Several safeguards can mitigate these risks. First, manual expert curation remains essential. Researchers should review generated materials for quality and theoretical alignment (Amaya et al. 2020; Daikeler et al. 2025; Klamm et al. 2025; Schwitter 2025). In this context, it is negligible whether the material was collected from natural sources, such as social media posts, or generated synthetically. Data quality aspects (Weiß et al. 2025) and responsibility are essential characteristics to consider when working with text as data. Second, stability testing helps assess consistency:

[8] Continuing along this path requires detailed sociodemographic information of the potential survey participants to be simulated. An example that integrates this into a consistent framework is Talking-to-Machines: https://github.com/talking-to-machines/talking-to-machines

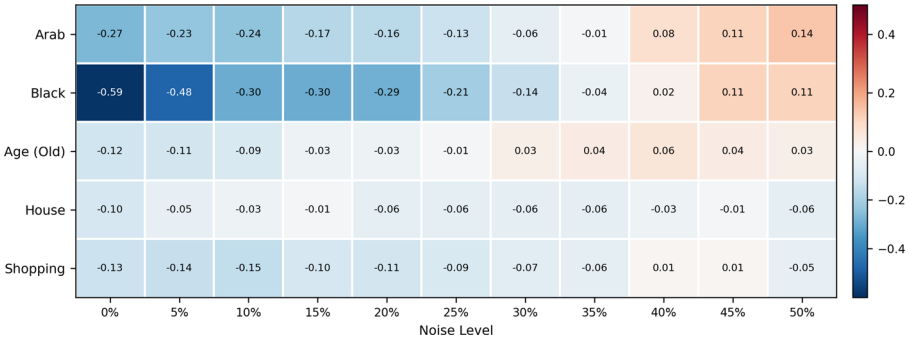

Fig. 7 Changes in regression coefficients under increasing levels of simulated label noise in the noise-exposure variable. The figure shows how the estimated associations between demographic variables and noise-exposure levels vary as the proportion of mislabelled observations increases (i.e. as simulated inter-annotator agreement decreases), illustrating the sensitivity of regression coefficients to different levels of label noise

Generating multiple versions with the same prompts and measuring semantic or visual similarity reveals whether outputs are reliable. As with stimulus sampling, these different versions can then also be used in surveys (Fong and Grimmer 2023). Third, where feasible, comparing results across different generation models or including some authentic materials alongside synthetic ones can test robustness (Goel et al. 2025). Finally, transparent documentation (Bender and Friedman 2018; Mitchell et al. 2019; Gebru et al. 2021; Klamm et al. 2025) of all prompts, model versions, and generation parameters enables critical evaluation and replication attempts.

Annotation Quality and Downstream Inference. Whether text is coded manually or automatically, annotation uncertainty propagates through causal estimation. Recent research has tried to quantify these risks systematically. Baumann et al. (2025), Bisbee and Spirling (2025), and Egami et al. (2023) demonstrate that synthetic annotations or annotation noise lead to incorrect statistical conclusions, even with advanced models. Critically, high annotation accuracy does not guarantee valid inference tasks. These errors take multiple forms: false positives (detecting non-existent effects), false negatives (missing real effects), sign errors (significant, but wrong direction), and magnitude errors (correct direction but wrong size). Standard statistical corrections such as doubly supervised learning (Hector and Song 2020) reduce false positives but increase false negatives. Egami et al. (2023), Ratzlaff et al. (2024), and Baumann et al. (2025) found that just a small number of human annotations outperform sophisticated automated techniques. This suggests treating classification tools as complex instruments requiring validation, not black-box replacements for measurement. Where feasible, trained human coders should establish gold standards. *When automation is necessary for scale, researchers should invest in validation against human-coded subsamples and conduct extensive sensitivity analyses.* In our text-as-outcome application, we deliberately used human annotation with multiple trained coders to ensure high-quality measurement despite the time investment required.

Example: To illustrate how labelling noise can substantially affect the causal pipeline, we use the annotations from our first empirical application as a concrete example. By introducing variation into our assistants' annotations, we simulate the higher uncertainty that typically characterises coding processes, particularly when mapping unstructured text, such as open-ended survey responses, into predefined label categories (Fig. 7).

In this example, we observe that our empirical findings are affected by noise levels exceeding 5%. However, it is also evident that the effects related to the variable "Black" are more robust, even with a certain increase in noisy labels. This robustness is further supported by the presence of multiple labels from three annotators, which reduces the impact of noise in annotations.

Researchers' Degrees of Freedom in Latent Treatment Discovery. Methods such as SIBP separate treatment discovery from effect estimation but still involve substantial interpretive judgment. Researchers must choose model configurations (number of features, prior specifications, etc.), decide which patterns are substantively meaningful, and assign labels to discovered treatments. Statistical formalism can obscure this interpretive work: A model may identify word patterns that predict outcomes, but deciding whether patterns represent "emotional exhaustion" versus "systemic critique" requires theoretical interpretation. Best practices include involving *multiple researchers in labelling* with formal IAA (coder reliability) assessment (Krippendorff 2011; McHugh 2012), pre-specifying theoretical expectations about plausible treatments, testing robustness across model specifications, and transparently reporting the decision path from raw output to final labels. *The goal is not to eliminate interpretation, intrinsic to social science, but to make it systematic and visible.*

Robustness Through Sensitivity Analysis. Given these uncertainties, robustness checks become essential. Rather than attempting to "correct" imperfect measurements, researchers should test how conclusions change under plausible measurement variations. This includes re-coding subsamples with different annotators or models, varying classification thresholds, or simulating measurement error at different rates to identify when conclusions would reverse. Importantly, Bisbee and Spirling (2025) show that models that are more accurate typically strengthen rather than weaken findings, indicating that significant results often become more robust as measurement improves. This suggests that researchers using reasonable but imperfect measures may commit more type II errors (false negatives/missing effects) than type I errors (false positives). *Null findings should therefore be interpreted cautiously as potentially reflecting measurement limitations.*

Practical Recommendations. The appropriate balance between automation and manual inspection depends on research goals and resources. However, several principles apply broadly:

- *Prioritise human judgment for construct validation:* Expert review of samples ensures that automated measures capture theoretically relevant variation.

- *Document all methodological choices:* Models, prompts, pre-processing, and configurations should be transparent and, ideally, be pre-registered.
- *Validate at every stage:* Assess generated text quality, compare automated and human coding, and verify treatment interpretability.
- *Conduct sensitivity analyses as standard practice:* Test whether conclusions depend on specific choices.
- *Acknowledge uncertainty:* Achieving perfect measurement is impossible; instead, the aim is to understand the potential size of the error and ensure robustness.

Text-as-data methods offer tremendous potential for sociology, but *realising this requires systematically addressing measurement challenges*. By *making uncertainty visible* rather than obscuring it, researchers can build cumulative knowledge about when textual data reliably support causal claims.

6 Conclusion

In this paper, we discussed and showcased approaches to causal inference with text as data. These approaches complement conventional sociological designs that rely on fixed, structured variables by exploiting the richness and abundance of textual data. Because text is often generated as a by-product of social processes and encodes beliefs, reasoning, and context, it can provide leverage on intermediate processes linking causes to outcomes. Across our applications, we highlight a central insight: Text comes not only with opportunities but also with several challenges. The sociological variables of interest—normative judgments, themes surrounding invisible labour—are never directly observable. Texts provide signals that require interpretation, whether through human coding, probabilistic feature models, or computational embeddings. Placed in the context of this special issue, our contribution complements and extends the broader conversation about how to best integrate and make use of new methodological developments in sociology.

What is the relationship between text-based causal inference and sociological explanation? We argue that text-based estimates are most informative for explanatory claims when (i) textual representations correspond to theoretically meaningful constructs, (ii) the temporal ordering between treatment, text, and outcome is well defined, and (iii) representational choices, such as codebook or embedding models, are transparent and validated out of sample. Under these conditions, variation in text (whether induced by experimental manipulation or observed as an outcome) can be interpreted as evidence about actor-level states or interaction processes such as shifts in attention, reasoning, or framing that are specified as part of a broader mechanism linking structural conditions to observable outcomes. Text, in this sense, provides empirical access to components of mechanisms that must be theoretically specified and discussed. Conversely, when these conditions are not met, text-based analyses should be understood primarily as descriptive (on descriptive analyses in sociology, see Goldthorpe 2026, this issue; Schwitter and Liebe 2025). Additionally, as discussed by Leitgöb and Keusch (2026), there are several promises and pitfalls of using large-scale data for sociological explanation surrounding questions about se-

lection, representation, and context that we have also encountered with text (see also Weiß et al. 2025). Text data originate from different sources, such as open-ended survey responses, administrative documents, and digital trace data such as social media posts. These texts are generated within specific social processes, institutional settings, and political contexts, and they reflect patterns of selection, expression, and communication. Meaningful causal analysis with text, therefore, requires attention to the data-generating process as well as theoretical expert knowledge.

We want to highlight that aspects of natural language processing, such as coding, feature extraction, and computational modelling, are never theory-neutral (see also Monroe et al. 2015). Choices about what to measure and how to measure it necessarily embed theoretical assumptions. For instance, deciding which words or topics matter reflects a prior conceptualisation of the phenomenon: In our text-as-outcome study, the decision to focus on criminalising-normative labels was guided by theory and linked further to an understanding of how norms are expressed in (German) language. Similarly, decisions on the unit of analysis—whether paragraphs, sentences, or whole documents are treated as units—reflect assumptions about social action and meaning. It was a deliberate choice how to design the synthetic Reddit posts in our text-as-treatment outcome, while facing a trade-off of the perspective-taking theory—which generally suggests more in-depth engagement with the other perspective—and survey methodological constraints.

This also relates back to the interpretive leeway inherent in probabilistic latent feature models, where transparency becomes essential, as we argued in the last section. New approaches (e.g. Imai and Nakamura 2024, 2025) offer exciting possibilities for latent treatment discovery and effect estimation, yet also amplify interpretive discretion (for a broader discussion of ML approaches, see also Jeon and Brand 2026, this issue). Recent work cautions that LLMs can be "prompt-hacked" to produce desired results, analogous to *p*-hacking in traditional analyses (Kosch and Feger 2025).

In this special issue, Hurtado Bodell et al. (2026) discuss the close connection between methodological innovation and substantive theorising, showing how text-as-data methods can actively contribute to explanation. In contrast to our small-scale corpora, they highlight how text data can also be used for causal inference at scale. Beyond enabling causal estimation, text-as-data approaches can also support theory development. By examining how individuals narrate and make sense of social phenomena, researchers can gain insight into latent mechanisms and causal processes that are often difficult to capture with traditional models or structured survey instruments. For instance, in our text-as-outcome study, several respondents' accounts reflect a dual-process dynamic: Initial reactions appeared to be shaped by ethnic stereotypes, which were subsequently reconsidered or overridden upon reflection.

Taken together, our article aimed at highlighting the challenges of explanation and causal inference in sociology when using text as data. Texts encode rich, context-dependent social information, yet extracting causal insight requires careful theoretical framing, rigorous methodological design, and transparent interpretive decisions.

Supplementary Information The online version of this article (https://doi.org/10.1007/s11577-026-01071-y) contains supplementary material, which is available to authorized users.

Acknowledgements We want to thank the editors of this special issue and the participants in the contributors' workshop for their helpful comments. We also want to thank Leonard Bek and Ella Häcker for their support in text labelling.

Funding The authors disclosed receipt of the following financial support for the research of this article: This work was supported by the Postdoc Career Academy of the University of Mannheim and the German Research Foundation's Emmy-Noether Program (ZH 613/1-1).

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability The text-as-outcome application uses data from wave 76 of the German Internet Panel (GIP). The text data used are only available through the on-site data access at the University of Mannheim. Labelled data can be made available upon request to researchers which have access to the GIP data. The data pertaining to the text-as-treatment application can befound here: https://doi.org/10.7910/DVN/SEV0SZ. All code is available on Github: https://github.com/rubac/text_treat.

Conflict of interest N. Schwitter, R.L. Bach, and C. Klamm declare that they have no competing interests.

References

Aher, Gati, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th international conference on machine learning (ICML '23)*. New York: Association for Computing Machinery (ACM).

Amaya, Ashley, Paul P. Biemer, and David Kinyon. 2020. Total error in a big data world: Adapting the TSE framework to big data. *Journal of Survey Statistics and Methodology* 8(1): 89–119.

Amaya, Ashley, Ruben Bach, Florian Keusch, and Frauke Kreuter. 2021. New data sources in social science research: Things to know before working with Reddit data. *Social Science Computer Review* 39(5): 943–960.

Argyle, Lisa P., Ethan C. Busby, et al. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis* 31(3): 327–343.

Barrie, Christopher, Alexis Palmer, and Arthur Spirling. 2025. Replication for language models: Problems, principles, and best practice for political science. *Working paper*.

Baumann, Joachim, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor M. Plaza-del-Arco, Johannes B. Gruber, and Dirk Hovy. 2025. Large language model hacking: Quantifying the hidden risks of using LLMs for text annotation. *arXiv*: 2509.08825 [cs.CL].

Beck, Tilman, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th conference of the European chapter of the association for computational linguistics (volume 1: long papers)*. Stroudsburg, PA: Association for Computational Linguistics.

Bender, Emily M. and Batya Friedman. 2018. Data statements for natural language processing: toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6: 587–604.

Benoit, Ken. 2020. Text as data: An overview. In *The SAGE handbook of research methods in political science and international relations*, 461–497. London: SAGE Publications Ltd.

Bisbee, James and Arthur Spirling. 2025. What to do when humans are no longer the old standard: Large language models, state of the art and robustness. Draft dated September 23, 2025. https://github.

com/ArthurSpirling/futureProofR/blob/main/Bisbee_Spirling_human_gold_standard_9_23_2025.pdf. Accessed 05.02.2026.
Bleich, Erik, Stonebraker Hannah, Nisar Hasher, and Rana Abdelhamid. 2015. Media portrayals of minorities: Muslims in British newspaper headlines, 2001–2012. *Journal of Ethnic and Migration Studies* 41(6): 942–962.
Blom, Annelies G., Christina Gathmann, and Ulrich Krieger. 2015. Setting up an online panel representative of the general population: The German internet panel. *Field Methods* 27(4): 391–408.
Breen, Richard, and Guanghui Pan. 2026. Bringing external validity into sociological research. *This issue.*
Broockman, David E., and Joshua L. Kalla. 2016. Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science* 352(6282): 220–224.
Bull, Raymond H. C., and J. Green. 1980. The relationship between physical appearance and criminality. *Medicine, Science and the Law* 20(2): 79–83.
Correll, Joshua, Bernadette Park, Charles M. Judd, and Bernd Wittenbrink. 2002. The police officer's dilemma: using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology* 83(6): 1314–1329.
Daikeler, Jessica, Leon Fröhling, Indira Sen, Lukas Birkenmaier, Tobias Gummer, Jan Schwalbach, Henning Silber, Bernd Weiß, Katrin Weller, and Clemens Lechner. 2025. Assessing data quality in the age of digital social research: A systematic review. *Social Science Computer Review* 43(5): 943–979.
Daminger, Allison. 2019. The cognitive dimension of household labor. *American Sociological Review* 84(4): 609–633.
Diekmann, Andreas. 2026. Analytical sociology. Generative models, mechanisms, explanations. *This issue.*
Duncan, Birt L. 1976. Differential social perception and attribution of intergroup violence: Testing the lower limits of stereotyping of Blacks. *Journal of Personality and Social Psychology* 34(4): 590–598.
Eagly, Alice H. 2013. *Sex differences in social behavior: A social-role interpretation*. Hove: Psychology Press.
Eberhardt, Jennifer L., Phillip A. Goff, Valerie J. Purdie, and Paul G. Davies. 2004. Seeing black: Race, crime, and visual processing. *Journal of Personality and Social Psychology* 87(6): 876–893.
Egami, Naoki, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2022. How to make causal inferences using texts. *Science Advances* 8(42): eabg2652.
Egami, Naoki, Musashi Hinck, Brandon M. Stewart, and Hanying Wei. 2023. Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. In *Advances in neural information processing systems*, ed. Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, 68589–68601. Red Hook: Curran Associates, Inc.
Feder, Amir, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics* 10: 1138–1158.
Fernandez, Jared, Clara Na, Vashisth Tiwari, Yonatan Bisk, Sasha Luccioni, and Emma Strubell. 2025. Energy considerations of large language model inference and efficiency optimizations. In *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: long papers)*, 32556–32569. Vienna: Association for Computational Linguistics.
Fong, Christian. 2019. texteffect: Discovering latent treatments in text corpora and estimating their causal effects. R package version 0.3.
Fong, Christian and Justin Grimmer. 2016. Discovery of treatments from text corpora. In *Proceedings of the 54th annual meeting of the association for computational linguistics,* 1600–1609. Berlin, Germany: Association for Computational Linguistics.
Fong, Christian and Justin Grimmer. 2023. Causal inference with latent treatments. *American Journal of Political Science* 67(2): 374–389.
Galinsky, Adam D. and Gordon B. Moskowitz. 2000. Perspective-taking: decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology* 78(4): 708–724.
Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer W. Vaughan, Hanna Wallach, Hal D. III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM* 64(12): 86–92.
Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. Text as data. *Journal of Economic Literature* 57(3): 535–574.
Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* 120(30): e2305016120.

Goel, Shashwat, Joschka Struber, Ilze A. Auzina, Karuna K. Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge, and Jonas Geiping. 2025. Great models think alike and this undermines AI oversight. *arXiv*: 2502.04313 [cs.LG].

Golder, Scott A., and Michael W. Macy. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* 333(6051): 1878–1881.

Goldthorpe, John H. 2026. Description, causal explanation and policy intervention in sociology. *This issue.*

Grimmer, Justin, and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21(3): 267–297.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2022. *Text as data: A new framework for machine learning and the social sciences*. Princeton (N.J.): Princeton University Press.

Haines, Elizabeth L., and Steven J. Stroessner. 2019. The role prioritization model: How communal men and agentic women can (sometimes) have it all. *Social and Personality Psychology Compass* 13(12): e12504.

Hainmueller, Jens, Daniel J. Hopkins, and Teppei Yamamoto. 2014. Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political Analysis* 22(1): 1–30.

Halterman, Andrew, and Katherine A. Keith. 2025. Codebook LLMs: Evaluating LLMs as measurement tools for political science concepts. *Political Analysis*: 1–17.

Hassan, Tarek A., Stephan Hollander, Aakash Kalyani, Laurence van Lent, Markus Schwedeler, and Ahmed Tahoun. 2025. Text as data in economic analysis. *Journal of Economic Perspectives* 39(3): 193–220.

Haupt, Andreas, and Dafna Gelbgiser. 2024. The gendered division of cognitive household labor, mental load, and family–work conflict in European countries. *European Societies* 26(3): 828–854.

Hector, Emily C., and Peter X. Song. 2020. Doubly distributed supervised learning and inference with high-dimensional correlated outcomes. *Journal of Machine Learning Research* 21(173): 1–35.

Hochschild, Arlie, and Anne Machung. 1989. *The second shift: Working parents and the revolution at home.* New York: Viking.

Holland, Paul W. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81(396): 945–960.

Hu, Tiancheng, and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd annual meeting of the association for computational linguistics*, 10289–10307. Bangkok, Thailand: Association for Computational Linguistics.

Huang, Yue, Zhengqing Yuan, Yujun Zhou, et al. 2024. Social science meets LLMs: How reliable are large language models in social simulations? *arXiv:*2410.23426 [cs.CL].

Hurtado Bodell, Miriam, Marc Keuschnigg, Ana Macanovic, and Anastasia Menshikova. 2026. Computational text analysis for building and testing social theory. *This issue.*

Imai, Kosuke, and Kentaro Nakamura. 2024. Causal representation learning with generative artificial intelligence: Application to texts as treatments. *arXiv*: 2410.00903 [stat.AP].

Imai, Kosuke, and Kentaro Nakamura. 2025. GenAI-powered inference. *arXiv*: 2507.03897 [cs.LG].

Jeon, Nanum, and Jennie E. Brand. 2026. Causal machine learning: A deductive-inductive framework for sociological research. *This issue.*

Jurafsky, Daniel, and James H. Martin. 2009. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* Upper Saddle River, NJ: Pearson.

Kahneman, Daniel. 2011. *Thinking, fast and slow.* New York: Farrar, Straus and Giroux.

Kaiser, Carolin, Jakob Kaiser, Vladimir Manewitsch, Lea Rau, and Rene Schallner. 2025. Simulating human opinions with large language models: Opportunities and challenges for personalized survey data modeling. In *Adjunct proceedings of the 33rd ACM conference on user modeling, adaptation and personalization*. New York, NY: Association for Computing Machinery.

Kalla, Joshua L., and David E. Broockman. 2023. Which narrative strategies durably reduce prejudice? Evidence from field and survey experiments supporting the efficacy of perspective-getting. *American Journal of Political Science* 67(1): 185–204.

Keuschnigg, Marc, Niclas Lovsjö, and Peter Hedström. 2018. Analytical sociology and computational social science. *Journal of Computational Social Science* 1(1): 3–14.

Klamm, Christopher, Ruben Bach, and Tornike Tsereteli. 2025. CARING: Enhancing open data quality through community engagement. datasets.cool.

Kosch, Thomas, and Sebastian Feger. 2025. Prompt-hacking: The new p-hacking? *arXiv*: 2502.14571 [cs.HC].

Krippendorff, Klaus. 2011. Computing Krippendorff's alpha-reliability. https://www.asc.upenn.edu/sites/default/files/2021-03/Computing%20Krippendorff%27s%20Alpha-Reliability.pdf. Last accessed: 05.02.2026.

Leitgöb, Heinz, and Florian Keusch. 2026. Causal inferences from digital behavioral data: Methodological implications. *This issue.*

Liu, Andy, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. In: *Findings of the association for computational linguistics: ACL 2024*. Bangkok, Thailand: Association for Computational Linguistics.

Luccioni, Sasha, Yacine Jernite, and Emma Strubell. 2024. Power hungry processing: Watts driving the cost of AI deployment? In *The 2024 ACM conference on fairness, accountability, and transparency*, 85–99. New York: Association for Computing Machinery (ACM).

Luccioni, Sasha, Boris Gamazaychikov, Sara Hooker, 2024. Light bulbs have energy ratings—so why can't AI chatbots? *Nature* 632(8026): 736–738.

MacLin, M K., and Vivian Herrera. 2006. The criminal stereotype. *North American Journal of Psychology* 8(2): 197–208.

Mayring, Philipp. 2015. *Qualitative Inhaltsanalyse: Grundlagen und Techniken*. 12th ed. Weinheim and Basel: Beltz.

McHugh, Mary. 2012. Interrater reliability: The kappa statistic. *Biochemia medica* 22: 276–282.

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa D. Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency.* FAT* '19, 220–229. New York: Association for Computing Machinery (ACM).

Modarressi, Iman, Jann Spiess, and Amar Venugopal. 2025. Causal inference on outcomes learned from text. *arXiv*: 2503.00725 [econ.EM].

Monroe, Burt L., Jennifer Pan, Margaret E. Roberts, Maya Sen, and Betsy Sinclair. 2015. No! Formal theory, causal inference, and big data are not contradictory trends in political science. *PS: Political Science & Politics* 48(1): 71–74.

Payne, B K. 2001. Prejudice and perception: the role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology* 81(2): 181–192.

Ratzlaff, Neale, Matthew L. Olson, Musashi Hinck, Shao-Yen Tseng, Vasudev Lal, and Phillip Howard. 2024. Debiasing large vision-language models by ablating protected attribute representations. *arXiv*: 2410.13976 [cs.CV].

Raub, Werner. 2026. Explanation in sociology and Coleman's diagram. *This issue.*

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana K. Gadarian, Bethany Albertson, and David G. Rand. 2014. Structural topic models for open-ended survey responses. *American Journal of Political Science* 58(4): 1064–1082.

Roberts, Margaret E., Brandon M. Stewart, and Richard A. Nielsen. 2020. Adjusting for confounding with text matching. *American Journal of Political Science* 64(4): 887–903.

Röttger, Paul, Bertie Vidgen, Dirk Hovy, and Janet B. Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 conference of the north American chapter of the association for computational linguistics: Human language technologies.* New York: Association for Computational Linguistics.

Rubin, Donald B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5): 688–701.

Rubin, Donald B. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association* 100(469): 322–331.

Santurkar, Shibani, Esin Durmus, Faisal Ladhak, Mina Lee, Percy Liang, and Tatsunori B. Hashimoto. 2023. Whose opinions do language models reflect? *Stanford institute for human-centered artificial intelligence (HAI) policy brief.* Stanford, CA: Stanford University.

Schonlau, Matthias, Julia Weiß, and Jan Marquard. 2023. Multi-label classification of open-ended questions with BERT. arXiv: 2304.02945 [stat.AP].

Schwitter, Nicole. 2025. Using artificial intelligence to generate visual vignettes in factorial survey experiments. *Social Science Computer Review:* 08944393251392916.

Schwitter, Nicole, and Ulf Liebe. 2025. Sometimes, a descriptive figure is worth more than a thousand model coefficients: the importance of data description in social research. *International Journal of Social Research Methodology*: 1–6.

Stelter, Marleen, Iniobong Essien, Anette Rohmann, Juliane Degner, and Stefanie Kemme. 2023. Shooter biases and stereotypes among police and civilians. *Acta Psychologica* 232: 103820.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 3645–3650. Florence: Association for Computational Linguistics.

Sun, Huaman, Jiaxin Pei, Minje Choi, and David Jurgens. 2025. Sociodemographic prompting is not yet an effective approach for simulating subjective judgments with LLMs. In *Proceedings of the 2025 conference of the nations of the Americas chapter of the association for computational linguistics,* 845–854. Albuquerque, NM: Association for Computational Linguistics.

Törnberg, Petter. 2023. How to use LLMs for text analysis. *arXiv:* 2307.13106 [cs.AI].

Treischl, Edgar, and Tobias Wolbring. 2022. The past, present and future of factorial survey experiments: A review for the social sciences. *Methods, data, analyses: A Journal for Quantitative Methods and Survey Methodology* 16(2): 141–170.

Tutic, Andreas, Sascha Grehl, and Ulf Liebe. 2024. A dual-process perspective on the relationship between implicit attitudes and discriminatory behaviour. *European Sociological Review* 40(4): 672–685.

Uma, Alexandra N., Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research* 72: 1385–1470.

Von der Heyde, Leah, Anna-Carolina Haensch, Bernd Weiß, and Jessica Daikeler. 2025. Using large language models for coding German open-ended survey responses on survey motivation. *Survey Research Methods* 19(4): 355–370.

Weber-Genzel, Leon, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)*, 2256–2269. Bangkok, Thailand: Association for Computational Linguistics.

Weiß, Bernd, Heinz Leitgöb, and Claudia Wagner. 2025. Conceptualizing, assessing, and improving the quality of digital behavioral data. *Social Science Computer Review* 43(5): 927–942.

Wood-Doughty, Zach, Ilya Shpitser, and Mark Dredze. 2018. Challenges of using text classifiers for causal inference. In: *Proceedings of the 2018 conference on empirical methods in natural language processing,* ed. Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, 4586–4598. Brussels, Belgium: Association for Computational Linguistics.

Xu, Yinuo, Veronica Derricks, Allison Earl, and David Jurgens. 2025. Modeling annotator disagreement with demographic-aware experts and synthetic perspectives. *arXiv:* 2508.02853 [cs.CL].

Yang, Zhimi, and Bo Shen. 2025. Estimating textual treatment effect via causal disentangled representation learning. *The Journal of Supercomputing* 81(2): 386.

Zhang, Bo, and Jiayao Zhang. 2022. Some reflections on drawing causal inference using textual data: Parallels between human subjects and organized texts. In *Proceedings of the first conference on causal learning and reasoning*, 1026–1036. Proceedings of Machine Learning Research.

Nicole Schwitter 1992, PhD, is a Postdoctoral Research Fellow at the Mannheim Centre for European Social Research (MZES), University of Mannheim, and an Honorary Research Fellow at the University of Warwick. Her research combines computational social science approaches with substantive interests in group relations, including discrimination, integration, and offline and online social relationships. Publications: Offline connections, online votes: The role of offline ties in an online public election. New Media & Society, 2024; Explaining ethnic violence: on the relevance of geographic, social, economic, and political factors in hate crimes on refugees. European Sociological Review, 2021 (with U. Liebe).

Ruben L. Bach 1989, Dr. rer. soc., is a Senior Research Fellow at the Mannheim Centre for European Social Research (MZES), University of Mannheim. Research areas: Computational social science, survey methodology, quantitative methods, digital trace data, gendered household and care work, and algorithmic decision-making. Publications: When Small Decisions Have Big Impact: Fairness Implications of Algorithmic Profiling Schemes. ACM Journal on Responsible Computing, 2024 (with C. Kern, H. Mautner, and F. Kreuter); Understanding political news media consumption with digital trace data and natural language processing. Journal of the Royal Statistical Society: Series A, 2022 (with C. Kern, D. Bonnay, and L. Kalaora).

Christopher Klamm 1990, is an interdisciplinary researcher with a background in Computer Science and Political Science (Darmstadt and Zurich). He is pursuing a PhD at the University of Mannheim, serving as a senior researcher at the Cologne Center of Comparative Politics, and is currently a research associate at Oxford's Computational Political Science Group. His research combines computational social science and traditional social science methods. Publications: AI Conversational Interviewing: Transforming Surveys with LLMs as Adaptive Interviewers. LaTeCH-CLfL 2025 (with A. Wuttke, M. Aßenmacher, M. M. Lang, Q. Würschinger and F. Kreuter); Our kind of people? Detecting populist references in political debates. EACL 2023 (with I. Rehbein and S. P. Ponzetto).