How does partisan type influence affective polarization?*

A comparative study of 25 European democracies

Tristan Muno†

December 10, 2025

Research agrees that affective polarization (AP) is pervasive but disagrees about why: identity-centered theories claim explicit partisan attachment primarily drives hostility, while substantive-inference accounts argue that partisan cues trigger expectations about ideology and social traits that can generate the same behaviors without felt identity. This distinction matters most in multiparty systems, where the share of explicit identifiers varies cross-nationally and measurement choices reshape the population of inference. We introduce partisan type—explicit versus implicit partisans with vote anchors—to isolate whether subjective attachment amplifies behavioral AP. Reanalyzing a large conjoint experiment in 25 European democracies (N 29,800) with dictator and trust games, we estimate ingroup favoritism and outgroup derogation across types. Across tasks and countries, explicit partisans discriminate no more than implicit ones. AP in Europe appears to arise primarily from categorization and inferred alignment, not from self-reported partisan identity, reshaping current debates over the origins and measurement of AP.

1 Introduction

Affective polarization (AP) has become a defining feature of political conflict across contemporary democracies (Iyengar et al. 2019; Boxell, Gentzkow, and Shapiro 2024; Reiljan et al. 2024). Beyond ordinary disagreement, AP entails social distancing, interpersonal hostility, and discriminatory behavior toward partisan outgroups (Iyengar, Sood, and Lelkes 2012; Iyengar and Wagner 2025), raising concerns about democratic cooperation

*Do not share without consent. Wordcount: 12623.

†University of Mannheim; Mail: tristan.muno@uni-mannheim.de

and political stability (Graham and Svolik 2020; Kingzette et al. 2021; Lars Eric Berntzen, Kelsall, and Harteveld 2023). Although scholars widely agree that AP is empirically pervasive, they disagree fundamentally about its source. Identity-centered accounts posit that partisan hostility reflects the unique psychological force of explicit, subjectively felt partisan attachment — the *party-over-policy* view (Huddy, Mason, and Aarøe 2015; Mason and Wronski 2018; Iyengar and Westwood 2015; N. Dias and Lelkes 2022). In contrast, substantive-inference accounts argue that partisan labels convey ideological, normative, and social information that can generate the same behavioral responses even in the absence of subjective attachment — the *policy-over-party* view (Orr, Fowler, and Huber 2023; Orr and Huber 2020; Webster and Abramowitz 2017). Thus, while AP itself is not in doubt, its core mechanism remains contested.

A central obstacle to adjudicating these perspectives is that identity-based and inference-based pathways are observationally equivalent (Orr, Fowler, and Huber 2023). Party cues simultaneously signal group membership and ideological content, making it difficult to isolate whether behavioral AP reflects the psychological force of identity or cue-driven expectations about agreement (Dafoe, Zhang, and Caughey 2018). This problem is compounded by conceptual and measurement heterogeneity in the AP literature (Campos and Federico 2025). Behavioral indicators such as allocation or trust decisions may reflect categorization or expected policy alignment as much as they reflect identity (Orr, Fowler, and Huber 2023). As a result, empirical tests often conflate mechanisms, leading to divergent claims about the origins of AP.

These inferential challenges are amplified in European multiparty systems (Röllicke 2023). Unlike the U.S. case, explicit partisan attachment is far less prevalent and varies substantially across countries, surveys, and party systems. Restricting analyses to explicit identifiers therefore shifts the estimand, reduces the population of inference, and

undermines cross-national comparability. Existing measures differ precisely in whom they treat as "partisans," and these differences translate into substantively different portraits of polarization (Reiljan 2020; Wagner 2021). In contexts where many citizens express stable partisan preferences while denying subjective attachment, the distinction between identity and categorization becomes especially consequential.

We contribute to these debates by introducing a generalizable distinction between explicit and implicit partisans — what we call partisan type. Explicit partisans report a felt attachment to a party, whereas implicit partisans deny such attachment yet nevertheless reveal a partisan anchor through vote choice. This distinction builds directly on classic work on independents (Kinder and Kalmoe 2017) and leaners (Petrocik 2009) as well as research on implicit partisan identity (Elliott 2024; Theodoridis 2017), extending these insights to multiparty systems where existing categories do not travel cleanly. By separating subjective attachment from behavioral anchoring, partisan type isolates the theoretical quantity that prior work has not directly examined: whether self-reported identity enhances behavioral AP beyond categorization and cue-based inference alone. It also provides a transparent framework for specifying the population of inference in comparative AP research — allowing scholars to clearly communicate whether their analyses focus on explicit identifiers alone or on the broader set of citizens who possess a partisan anchor, regardless of whether they report attachment.

Empirically, we leverage this distinction in a cross-national reanalysis of a large conjoint experiment conducted in 25 European democracies, paired with dictator and trust games that cleanly measure discriminatory behavior. Within-respondent randomization of partisan cues identifies the causal effect of encountering an ingroup or outgroup label, while hierarchical Bayesian modeling estimates the observational contrast between explicit and implicit partisans across countries and behavioral settings.

Our results indicate a consistent pattern. Across allocative and reciprocal tasks and across all 25 countries, explicit partisans discriminate no more than implicit partisans. Implicit partisans — who deny feeling attached — exhibit the same levels of ingroup favoritism (IF) and outgroup derogation (OD) as explicit identifiers. These findings suggest that much behavioral AP in Europe arises from partisan categorization and cue-driven inference rather than from subjective attachment, complicating identity-first interpretations and strengthening policy-over-party explanations.

These findings carry important implications for the debates that motivate this study. First, for measurement and inference, our findings validate the practice of combining vote-anchored and identifier respondents when estimating behavioral AP: there is no empirical reason to exclude implicit partisans, and doing so risks unnecessary loss of generalizability. Second, for theories of partisan identity, the standard PID attachment item appears only weakly related to discriminatory behavior — behavioral AP in Europe seems to arise primarily from categorization, inferred ideology, and social expectations rather than from self-reported attachment. Third, the cross-national consistency of the null effect suggests a broadly shared mechanism of cue processing that operates across diverse party systems. Taken together, these results challenge identity-only accounts of AP and strengthen the case for multidimensional, cue-driven models of partisan behavior in comparative politics.

2 Theory

The Rise of Affective Polarization

With *Affect, Not Ideology*, Iyengar, Sood, and Lelkes (2012) shifted longstanding debates about polarization away from substantive policy disagreement (Poole and Rosen-

thal 1984; McCarty, Poole, and Rosenthal 2006; Fiorina, Abrams, and Pope 2006; Fiorina, Abrams, and Pope 2008; Fiorina and Abrams 2008; Abramowitz and Saunders 2006, 2008; Abramowitz 2010) toward a more social and emotional understanding of partisan conflict, emphasizing that partisans increasingly favor their own side while disliking or discriminating against the opposing party (Iyengar and Westwood 2015; Iyengar et al. 2019; Iyengar and Wagner 2025). This reconceptualization catalyzed a vast research program investigating the drivers of affective polarization (AP) (Iyengar et al. 2019; Hobolt, Lawall, and Tilley 2024; Bail et al. 2018; N. Dias and Lelkes 2022; Orr and Huber 2020; Rogowski and Sutherland 2016; Matthew S. Levendusky 2013; Mason 2015, 2016, 2018a; Törnberg 2022; Luttig 2017, 2018), its social, political, and economic consequences (Diermeier and Li 2019; Lars Eric Berntzen, Kelsall, and Harteveld 2023; McConnell et al. 2018, 2018; Jenke 2024; Kingzette et al. 2021; Huber and Malhotra 2017; Broockman, Kalla, and Westwood 2023; A. H.-Y. Lee 2022; S. Lee, Choi, and Ahn 2025; Graham and Svolik 2020), its temporal evolution (Iyengar, Sood, and Lelkes 2012; Iyengar and Krupenkin 2018; Iyengar et al. 2019), and possible interventions to reduce it (Adams et al. 2023; Kalla and Broockman 2022; Matthew S. Levendusky 2018; J. N. Druckman, Green, and Iyengar 2023; Mernyk et al. 2022). Initially focused on the American case, scholars have increasingly studied AP in Europe (Wagner 2021; Reiljan 2020; Hahm, Hilpert, and König 2024) and comparative contexts worldwide (Reiljan et al. 2024; Torcal, Reiljan, and Zanotti 2023; Ferreira da Silva and Garzia 2024; Garzia, Ferreira da Silva, and Maye 2023; Boxell, Gentzkow, and Shapiro 2024; Gidrom, Adams, and Horne 2020). Yet the rapid expansion of this literature, while demonstrating the broad reach of AP, has also generated conceptual and measurement ambiguities, as different operationalizations capture distinct facets of the phenomenon and yield divergent conclusions about its underlying nature (Campos and Federico 2025; Bakker and Lelkes 2024; Carlin and Love 2025; Iyengar

and Wagner 2025) This led scholars to refine theories of AP by clarifying who counts as partisan, distinguishing between horizontal hostility toward citizens and vertical hostility toward elites (Klar, Krupnikov, and Ryan 2018; Areal and Harteveld 2024; J. Druckman and Levendusky 2019), with the latter further differentiated into leader and party AP (Reiljan et al. 2024), and addressing the added complexity of multiparty systems where ingroup and outgroup boundaries are less symmetric (Röllicke 2023; Torcal and Comellas 2025).

Conceptual and Measurement Ambiguities

These debates intersect with normative concerns about which forms of AP pose tangible risks to democratic norms (Moore-Berg et al. 2020; Mernyk et al. 2022; Lars Eric Berntzen, Kelsall, and Harteveld 2023; Kingzette et al. 2021) and which fears may be overstated (Voelkel et al. 2023; Westwood et al. 2022a, 2022b). Most recently, Campos and Federico (2025) proposed a multidimensional framework conceptualizing AP as comprising interrelated dimensions of othering, aversion, and moralization (see also Finkel et al. 2020), accompanied by measurement strategies tailored to each subdimension. This refinement is normatively consequential because different facets of AP vary substantially in their links to democratic erosion and political violence (Lars Erik Berntzen 2025; Lars Eric Berntzen, Kelsall, and Harteveld 2023; Kalmoe and Mason 2022b; Moore-Berg et al. 2020). Thus, although there is broad consensus that AP is empirically widespread, its fundamental sources and normative implications remain contested (Westwood et al. 2022a; Kalmoe and Mason 2022a; Broockman, Kalla, and Westwood 2023), motivating the need for more precise distinctions in theory and measurement — including the role of different forms of partisanship, which is the focus of this paper.

Identity-Based Mechanisms of Partisan Behavior

Social identity theory (SIT) provides the dominant framework for interpreting AP in political contexts. Classic work shows that group identification generates predictable patterns of ingroup favoritism (IF) and outgroup derogation (OD) — even when group boundaries are minimal or arbitrary (Henri Tajfel et al. 1971; Henri Tajfel 1974; H. Tajfel and Turner 1979; Henri Tajfel and Turner 2004; Sherif et al. 1961; Allport 1954; Brewer 1999). Applied to politics, these mechanisms imply that stronger partisan attachment should heighten discriminatory behavior toward partisan targets (Huddy, Mason, and Aarøe 2015; Mason 2013; Iyengar and Westwood 2015). Contemporary accounts, such as the social-sorting perspective, further argue that when multiple social identities align with partisanship, partisan identity becomes more salient and conflictual, intensifying these behavioral tendencies (Greene 1999; Huddy 2013; Huddy, Mason, and Aarøe 2015; Huddy and Bankert 2017; Huddy, Bankert, and Davies 2018; West and Iyengar 2022; Mason 2013, 2015, 2016, 2018a, 2018b; Mason and Wronski 2018; Roccas and Brewer 2002).

Substantive Inference and the Limits of Identity-Only Explanations

Despite the prominence of identity-based explanations for AP, a growing body of work questions whether partisan identity alone drives discriminatory behavior, highlighting instead the role of ideological and substantive inferences (Bougher 2017; Webster and Abramowitz 2017; Orr and Huber 2020). When individuals encounter a party label, they routinely update beliefs about policy positions, group norms, and social composition, making it difficult to isolate the behavioral effects of identity per se from those of inferred agreement (Dafoe, Zhang, and Caughey 2018; Orr, Fowler, and Huber 2023). Whereas

the "party-over-policy" view posits that partisan identity dominates ideology in generating AP (N. Dias and Lelkes 2022; Iyengar et al. 2019), a contrasting "policy-over-party" perspective argues that perceived substantive (dis)alignment explains affective responses more than loyalty or identity strength (Bougher 2017; Webster and Abramowitz 2017; Orr and Huber 2020). Orr, Fowler, and Huber (2023) show that experimental manipulations intended to vary partisan identity simultaneously shift perceived policy agreement, implying that identity-based and policy-based pathways often produce observationally equivalent behaviors. This inferential entanglement extends even to behavioral experiments designed to cleanly measure IF and OD: partisan cues may function both as identity signals and as heuristics for ideology, and existing designs cannot fully disentangle the two mechanisms (Dafoe, Zhang, and Caughey 2018). Minimal-group and categorization research further complicates attribution by demonstrating that even arbitrary or shallow group boundaries can generate intergroup bias without subjective attachment (Sherif et al. 1961; Henri Tajfel 1974), suggesting that simply anchoring individuals to a party or vote choice may be sufficient to trigger partisan discrimination. Collectively, these insights crystallize a persistent "chicken-or-the-egg" problem: does partisan identity shape citizens' policy perceptions and preferences (N. C. Dias, Lelkes, and Pearl 2025), or do shared preferences and ideological affinities generate the sense of partisan identity that AP research often treats as fundamental (N. Dias and Lelkes 2022; Orr, Fowler, and Huber 2023)?

Comparative Challenges in Multiparty Systems

A further challenge arises when shifting from the U.S. two-party context to the multiparty systems that characterize most European democracies, where the structure of partisan competition renders the study of AP substantially more complex (Röllicke 2023; Wag-

ner 2021). Existing solutions vary widely: some approaches average evaluations across all out-parties, implicitly assuming a single ingroup (Reiljan 2020; Hahm, Hilpert, and König 2024), others measure the overall dispersion of like–dislike scores within a party system Wagner (2021), and still others incorporate the possibility of multiple ingroups or broader ideological blocs (Kekkonen and Ylä-Anttila 2021; Kekkonen et al. 2022; Bantel 2023; Reiljan and Ryan 2021). Yet these conceptual refinements underscore a deeper problem: measurement choices concerning who counts as a partisan — and thus who is included in the analysis — carry major implications for comparative inference (Garzia, Ferreira da Silva, and Maye 2023). In Europe, rates of self-reported partisan attachment vary dramatically across countries and surveys, meaning that restricting analyses to explicit identifiers changes both the estimand and the population of interest, thereby undermining cross-national comparability. Even the most widely used European AP measures — Reiljan (2020) AP index and Wagner (2021)'s spread measure — differ precisely in how they operationalize the in-party, with the former focusing on reported identifiers and the latter encompassing the whole electorate. Because most studies aim to draw conclusions about the population-level state or trajectory of AP, inference based solely on partisan subsamples is valid only if identifiers constitute a sufficiently large share of the electorate or accurately represent non-identifiers — conditions that rarely hold, given that the proportion of identifiers ranges from roughly 40% to 80% in our data. These complications are heightened in contexts where the "poles of polarization" are less clearly structured than in the American two-party system (Rothers 2025). Unsurprisingly, scholars increasingly emphasize that conceptualization and measurement design strongly shape substantive conclusions about AP, strengthening the case for developing more generalizable frameworks — such as partisan-type distinctions — that travel across party systems (Campos and Federico 2025; Carlin and Love 2025; Iyengar and Wagner 2025; Torcal and Comellas

2025)

A Generalizable Distinction: Explicit vs. Implicit Partisans

To address these conceptual and comparative limitations, we introduce a generalizable distinction between *explicit* and *implicit* partisans — what we term *partisan type*. Explicit partisans are individuals who report a felt attachment to a party, whereas implicit partisans are those who deny such attachment yet nevertheless express a partisan preference through loser-follow-up items or vote-choice questions. This definition captures the theoretically crucial distinction between respondents who affirm subjective attachment and those who explicitly deny it while still revealing a partisan anchor. Our approach synthesizes and extends classic insights from research on independents and leaners (Petrocik 2009; Elliott 2024; Kinder and Kalmoe 2017; Klar and Krupnikov 2016; Fowler et al. 2023) as well as on implicit partisan identity (Hawkins and Nosek 2012; Theodoridis 2017) to multiparty systems, thereby disentangling subjective identity from behavioral anchors in a way existing measures cannot. Although conceptually closest to the U.S. notion of leaners, we adopt the more general term partisan type because leaner is tied to a distinctly American partisan structure in which self-identified independents occupy a meaningful middle category between two poles. In multiparty Europe — where voters confront multiple ideological camps rather than a single axis — this metaphor does not travel cleanly, even though the underlying behavioral logic is similar. The theoretical utility of partisan type lies in its ability to adjudicate between identity-based and substance-based mechanisms of AP: if explicit partisans exhibit stronger discriminatory behavior than implicit partisans, this pattern would support identity-centered accounts, whereas similar behavior across types would indicate that partisan cues, ideological inferences, or categorization dynamics alone suffice to generate AP in the absence of reported

attachment (Orr, Fowler, and Huber 2023). To the best of our knowledge, no existing work has systematically examined whether behavioral manifestations of AP differ between explicit and implicit partisans. Conceptually, partisan type is narrower and more precise than "supporters" — a category that encompasses both explicit and implicit types — and avoids conflating subjective attachment with mere behavioral or expressive alignment. By providing a simple but theoretically granular distinction that can be applied across party systems, partisan type clarifies scholars' population of inference — whether the focus is on identifiers (explicit partisans *only*) or on the broader electorate (explicit *and* implicit partisans) — and thereby strengthens the analytical foundations for comparative research on AP.

Hypotheses: Behavioral Expectations Across Partisan Types

The distinction between explicit and implicit partisans provides a direct empirical test of two competing explanations for AP. The identity-centered account holds that subjective partisan attachment is a unique psychological force: explicit identifiers should react more strongly to partisan cues because identity heightens ingroup affect, perceived threat, and willingness to derogate outgroups. Under this view — the party-over-policy hypothesis — explicit partisanship should magnify all components of AP. In contrast, the substance/categorization account posits that partisan labels convey ideological, social, and normative information that suffices to generate discriminatory behavior even without felt attachment. This policy-over-party hypothesis predicts that implicit partisans, despite denying attachment, will discriminate to the same extent as explicit identifiers because categorization and inferred policy alignment — not subjective identity — drive affective and behavioral responses. Our contribution is to adjudicate directly between these two explanations by comparing explicit and implicit partisans in identical behavioral settings

across 25 multiparty systems.

Formally, these competing logics yield different hypotheses regarding the extent to which partisan types differ in their AP behaviors. If explicit attachment exerts an independent identity effect, then the party-over-policy hypothesis would predict:

- **H1a**: Explicit partisans exhibit *greater* AP than implicit partisans.
- **H2a**: Explicit partisans show *greater* IF than implicit partisans.
- **H3a**: Explicit partisans express *greater* OD than implicit partisans.

Conversely, if the policy-over-party mechanism holds and partisan cues and inferred ideological (dis)agreement would be sufficient to generate AP absent reported attachment, the following empirical implications should hold:

- **H1b**: Explicit and implicit partisans *do not* differ in AP.
- **H2b**: Explicit and implicit partisans *do not* differ in IF.
- **H3b**: Explicit and implicit partisans *do not* differ in OD.

Empirical Leverage and Theoretical Contribution

Together, these hypotheses allow us to distinguish whether behavioral AP in multiparty systems reflects the unique psychological force of explicit partisan identity (H1a–H3a) or whether partisan anchoring and cue-driven inferences alone are sufficient to reproduce AP's behavioral signature (H1b–H3b). Our empirical strategy is designed to leverage these competing mechanisms while remaining clear about the boundaries of what can be identified. By using a conjoint experiment with within-respondent randomization of partisan cues, we secure causal identification of how individuals respond to profiles labeled with ingroup or outgroup party information, independent of respondent characteristics.

We then pair this with hierarchical Bayesian models and g-computation to estimate the observational contrast between explicit and implicit partisans while adjusting for covariates and country-level clustering — precisely the combination required to adjudicate whether subjective attachment magnifies behavioral AP or whether partisan anchors alone suffice. Although our design cannot fully decompose the identity and substantive components embedded in the party cues themselves, it allows us to test a question that prior work has not: whether explicitly reported partisan attachment — often used as a proxy for identity — adds distinct behavioral content beyond what is already induced by partisan categorization and policy inference. In doing so, we contribute to the literature in three main ways. First, we introduce partisan type as a conceptual refinement that clarifies the population of inference in empirical AP research, extending the logic of U.S. "leaners" and implicit partisan identity to multiparty systems. Second, we document substantial cross-national variation in the prevalence of explicit and implicit types, underscoring the measurement heterogeneity that complicates comparative claims about identity-driven polarization. Third, we isolate the behavioral consequences of explicit attachment across two games and 25 countries, allowing us to evaluate directly whether explicit partisanship amplifies affective polarization. If explicit types exhibit stronger IF and OD, this would bolster identity-centered, party-over-policy accounts; if they behave similarly to implicit types, this would imply that reporting a felt attachment adds little beyond the informational and categorization processes captured by policy-over-party perspectives. In either case, our results speak directly to ongoing debates about the mechanisms underlying AP and refine both Orr, Fowler, and Huber (2023)'s critique and the multidimensional framework of Campos and Federico (2025) by identifying which behavioral components of AP are — and are not — meaningfully shaped by subjective partisan identity.

3 Research Design

In this section, we detail our empirical strategy for identifying the effect of partisan type on affective polarization (AP). We begin by introducing the dataset and experimental design, then describe the construction of partisan types and anchors, including a cross-national overview of their distribution. This distributional evidence clarifies the scope of the analytical sample and motivates our comparative research design. We then outline the key variables of interest and present a directed acyclical graph (DAG) that summarizes our causal model of the data-generating process. Next, we specify the statistical model and explain how we obtain the quantities of interest, followed by a justification of our covariate set. Finally, we enumerate the assumptions required for causal interpretation and situate our design within the broader literature, highlighting both strengths and remaining limitations.

3.1 Data and Experimental Design

To evaluate how different partisan types influence AP, we reanalyze the cross-national conjoint experiment conducted by Hahm, Hilpert, and König (2024). The study was fielded by Dynata (formerly SSI) in 25 EU member states between late May and mid-August 2019.[1] National samples consist of approximately 1,100 respondents per country (29,827 in total), drawn to be broadly representative on key sociodemographic dimensions. Summary statistics are shown in the appendix.

The experiment comprised two behavioral games: a dictator game capturing unilateral prosociality and a trust game introducing reciprocal incentives. As the games capture conceptually distinct behavioral settings, they will be analyzed separately. In each round,

[1]The countries covered are all EU member states (which included the UK at the time of data collection) except Luxembourg, Malta and Cyprus.

respondents interacted with a hypothetical partner with attributes ranging from age, gender, class, religion, to nationality and partisanship.[2] All conjoint attributes, including the partisan cue, were randomized independently with one restriction: partisan labels were displayed only when the profile's nationality matched the respondent's own country. This restriction was implemented to avoid implausible cross-national partisan combinations and maintain realism of the profile space. As a consequence, our analysis is restricted to the subset of profiles in which the partner is a conational. Within this restricted set, the control condition corresponds to conational profiles without a displayed partisan label. These restrictions do not affect randomization of the remaining attributes and preserve the "as-if random" variation in partisan cues among conational partners. Each respondent completed three rounds of each game, yielding six observations per participant. Our unit of analysis is thus a single game round. The resulting data are inherently nested: multiple rounds within respondents, respondents nested within party affiliations, and respondents and parties clustered within countries. We therefore rely on hierarchical models, described in detail in Section 3.3.

These behavioral measures offer several advantages for studying AP. First, the design is explicitly suited to assessing *horizontal* AP among mass partisans (Areal and Harteveld 2024). Common observational indicators — such as like/dislike scales or feeling thermometers — conflate horizontal evaluations of fellow citizens with *vertical* evaluations of party elites, thereby limiting their interpretability (J. Druckman and Levendusky 2019). Second, the inclusion of a nonpartisan baseline condition enables us to causally decompose AP into its constituent components: ingroup favoritism (IF) toward copartisans and outgroup derogation (OD) toward rival partisans. We estimate these components using standard conjoint estimands that capture the average marginal effect of each attribute on

[2]Detailed information, instruction wording and an exemplary profile are presented in the appendix.

token allocations (Hainmueller, Hopkins, and Yamamoto 2014).

Third, respondents hold multiple, potentially cross-cutting social identities (Roccas and Brewer 2002). The randomized conjoint design mitigates bias arising from such heterogeneity by independently varying several attributes that might otherwise be confounded with partisan cues. This feature allows us to isolate the specific effect of partisanship and reduces the risks of "aliasing" (Hainmueller, Hopkins, and Yamamoto 2014) and "masking" (Bansak et al. 2021), which occur when omitted attributes correlate with treatment dimensions (Dafoe, Zhang, and Caughey 2018).[3] Finally, by embedding partisan information among multiple randomized cues, conjoint tasks are less susceptible to social desirability pressures than direct survey questions (Horiuchi, Markovich, and Yamamoto 2022).

3.2 Measurement and Variable Construction

We construct our main variables of interest following a tripartite framework to ensure conceptual coherence and terminological consistency in all party-related measures. First, we define T_i as the *partisan type* of respondent i. If respondent i reported a party to which she feels attached to and is thus considered an *explicit* partisan, we define $T_i = 1$. Conversely, we set $T_i = 0$ for *implicit* partisans, i.e., respondent i who indicated no partisan attachment but reported a vote preference. We term the party reported by respondents *partisan anchor* A_i.[4]

[3]Naturally, a conjoint design can only include a limited set of attributes and therefore cannot preclude all forms of aliasing; we make no claim to exhaustiveness. Nonetheless, by randomizing several plausibly correlated characteristics (e.g., age, gender), the design reduces the risk that inferences about partisanship merely reflect respondents' assumptions about the social composition of party supporters. In this respect, it is substantially less susceptible to the confounding that affects like/dislike scales or feeling thermometers, which cannot disentangle the effect of partisanship from correlated beliefs about associated social groups.

[4]We use the term *partisan anchor* to denote the party reference point reported by respondents, regardless of whether they express subjective attachment to that party. For explicit partisans, this anchor corresponds to the conventional *partisan ingroup* in the literature (as in Hahm, Hilpert, and König (2024)). For implicit partisans, however, respondents explicitly deny a felt attachment, making social-

To evaluate the scope conditions of our design and to clarify the population of inference, we begin by documenting the distribution of partisan types in each of the 25 country samples. Figure 1 presents the share of respondents classified as explicit partisans, implicit partisans, and nonpartisans (type undefined). Two features of these descriptive patterns are substantively important. First, the proportion of nonpartisans is remarkably stable across countries (ranging mostly between 10 and 15%, least in Estonia (6%), most in Croatia (23%)), suggesting that the exclusion of this group — required for treatment consistency in our design — does not introduce systematic cross-national biases in the analytical sample. Second, and more consequential for our theoretical claims, the balance between explicit and implicit partisans varies dramatically across national contexts. In some electorates, explicit identifiers constitute the clear majority of partisans (e.g. strongest case being Denmark with 85% explicit partisans), in others, implicit partisans are equally prevalent (e.g. in Latvia both explicit and implicit partisans account for 45% of respondents). This heterogeneity underscores the need for a comparative framework capable of assessing whether implicit and explicit partisans differ systematically in their affective responses. In short, the figure motivates both the internal validity of our sample restriction and the broader comparative question that animates the study: whether distinct partisan types, whose prevalence differs markedly across countries, exhibit different patterns of AP.

In the experiment, each conjoint profile in round r contains a randomized party cue (Z_r^{party}). We compare this cue to the respondent's anchor to classify the *partisan relationship* R_{ri} as copartisan, outpartisan, or neutral ($R_{ri} \in \{Co, Out, None\}$). This coding corresponds to the average across outgroups operationalization (Reiljan 2020) and offers a tractable strategy for multi-party contexts, although it constitutes a substantial reduction

identity terminology inappropriate. The term *anchor* therefore provides a neutral label that applies consistently across both groups.

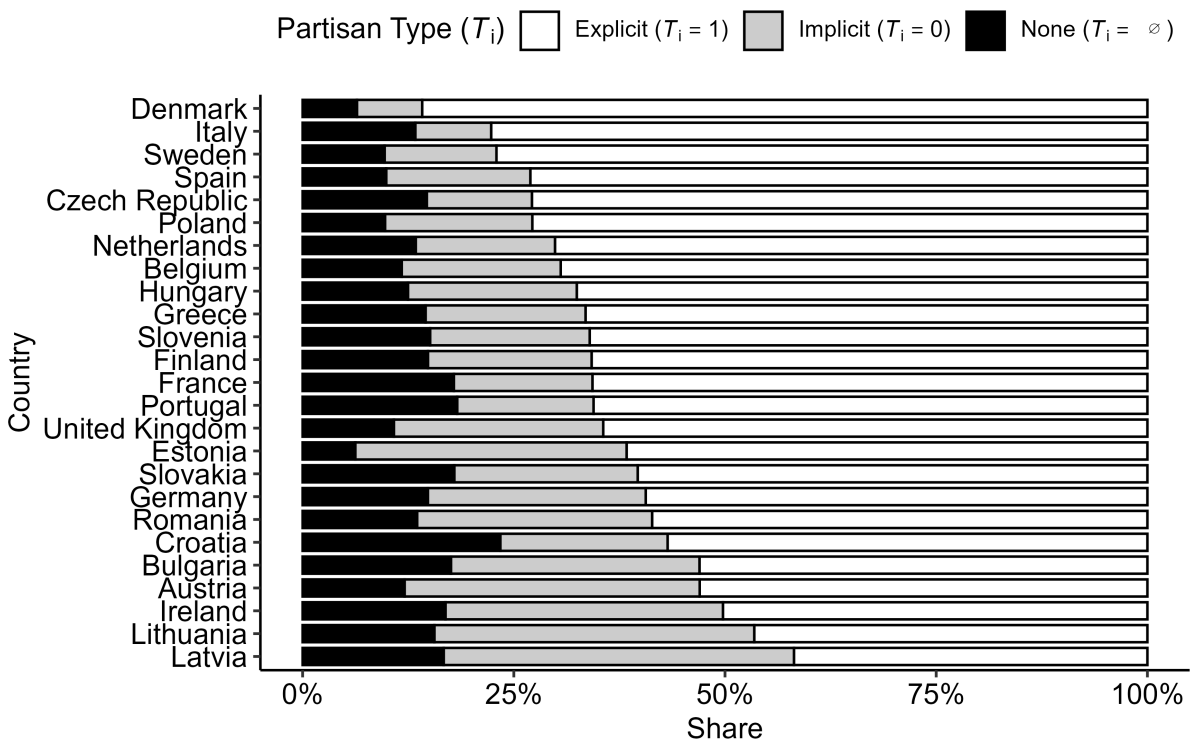

Figure 1: Distribution of partisan types, by country. Stacked horizontal bars show the within-country share (%) of three partisan types: explicit partisans (respondents who reported a subjective attachement to a party, $T_i = 1$), implicit partisans (respondents who reported no attachement but did report a vote preference or intention, $T_i = 0$), and respondents who reported neither (none, $T_i = \emptyset$). Percentages sum to 100% within each country, with country samples containing about $1,100$ respondents each (detailed numbers are reported in appendix section X).

of the inherent complexity of AP in multi-party systems (Röllicke 2023). We adopt this measure as it provides a uniform and comparable classification scheme across countries.

We conceptually frame explicit partisanship as the treatment and implicit partisanship as a control. Respondents reporting a subjective attachment to a political party ($T_i = 1$) thus form the treated group, whereas respondents expressing vote choice without attachment ($T_i = 0$) constitute the comparison group.[5] The underlying research question is: Does explicit partisan attachment amplify AP relative to implicit partisan attachment? In causal terms, we ask how a hypothetical intervention shifting T_i from 0 to 1 would change AP, holding other factors constant.

Figure 2 depicts the underlying causal data-generating process as a directed acyclical graph (DAG). As noted above, respondents report a partisan anchor A_i. Partisan type T_i is defined only for the subset of respondents who either explicitly or implicitly report a party and therefore appears as a child node of A_i. In each round r of the conjoint games, respondents receive a set of conjoint attributes, denoted by Z_r. The partisan cue Z_r^{party} and respondents' anchors A_i jointly determine the partisan relationship category via the mapping $f : (A_i, Z_r^{\text{party}} \mapsto R_{ij} \in \{\text{Co}, \text{Out}, \text{None}\})$.

The dependent variable Y_{riac} is the token allocation made in round r by respondent i with anchor a from country c. Token allocations Y are affected by both R_{ri} and Z_r^{other}, with the effect of R_{ri} theorized to be conditional on T_i. Because T_i is observational, we cannot assume conditional independence. We therefore include C_i to capture individual-level confounders affecting both T_i and Y_{riac}. Graphically, our substantive interest lies in the causal chain $T_i \rightarrow R_{ri} \rightarrow Yriac$, namely the effect of R_{ri} on Y_{riac} conditional on T_i, represented as the central vertical path in Figure 2.

[5] For readability, we omit the prefix "quasi-" when referring to treatment and control, without implying random assignement. Our design remains observational with respect to our treatment of interest, partisan type T.

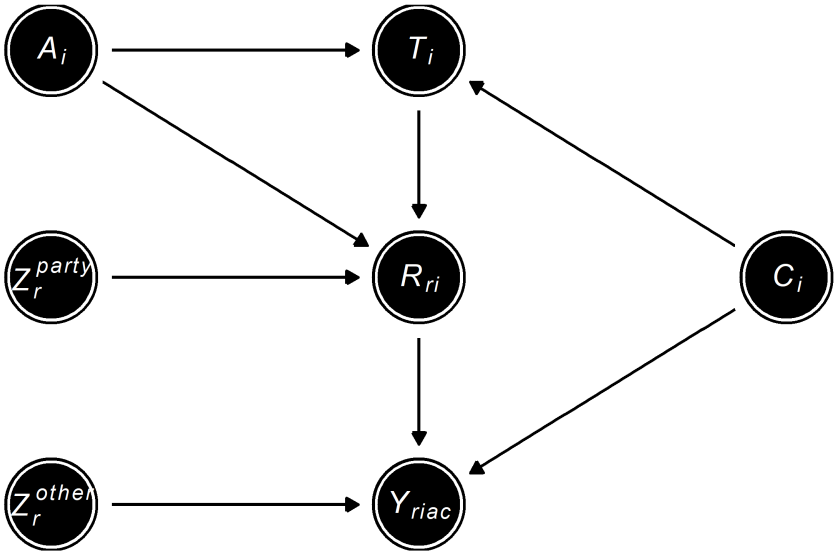

Figure 2: Directed acyclic graph of the causal data-generating process. Respondents have a partisan anchor A_i, representing the party they feel attached to (explicit partisans) or intend to vote for (implicit partisans). Partisan type T_i (explicit vs. implicit) is only defined for respondents with a partisan anchor and is therefore a child node of A_i. Each conjoint profile shown in round r presents randomized attributes: a partisan cue Z_r^{party} and other attributes Z_r^{other}. The partisan-relationship variable $R_{ri} = f(A_i, Z_r^{party})$ determines whether the profile is interpreted as a co-partisan, out-partisan, or neutral for respondent i. Token allocations Y_{riac} are affected by both R_{ri} and Z_r^{other}, with the effect of R_{ri} theorized to depend on T_i. Because T_i is observational, C_i denotes potential confounders of both T_i and Y_{ri}, highlighting the assumptions required for causal interpretation.

3.3 Model Specification and Estimation

We model the data in their natural hierarchical structure. The unit of analysis is a single conjoint profile evaluation in round r of a game. Profile evaluations (r) are nested within respondents (i), respondents are grouped by their partisan anchors (a), i.e., the party to which they report attachment or vote intention, and anchors are nested within countries (c). This yields a four-level structure: Level 1 (round r) $\rightarrow$ Level 2 (respondent i) $\rightarrow$ Level 3 (anchor a) $\rightarrow$ Level 4 (country c).

Let Y_{riac} denote the number of tokens allocated (0-10) in round r by respondent i with partisan anchor a in country c. Respondent partisan type is $T_i \in \{0, 1\}$ (implicit vs explicit). The respondent-profile partisan relationship $R_{ri} \in \{\text{Co, Out, None}\}$ is a deterministic function of the respondent's anchor and the profiles partisan cue. Z_r^{other} denotes the matrix of all other conjoint attributes, and C_i constitutes a matrix of respondent-level covariates discussed in Section 3.5.

To estimate how explicit partisan type conditions responsiveness to partisan cues, we specify a multilevel model with a cross-level interaction between T_i and R_{ri} and random effects at the respondent, anchor, and country levels, as shown in Equation 1:

$$\begin{aligned} Y_{riac} = {} & \beta_0 + \beta_1 T_i + \beta_2 R_{ri} + \beta_3 (T_i \times R_{ri}) + \beta_4 Z_{ri}^{\text{other}} + \beta_5 C_i \\ & + u_{i0} \\ & + v_{a0} + v_{aT} T_i + v_{aR} R_{ri} \\ & + w_{c0} + w_{cT} T_i + w_{cR} R_{ri} \\ & + \varepsilon_{riac} \end{aligned} \tag{1}$$

We assume a Gaussian likelihood for the bounded (0-10, see Figure 10 in the appendix) token allocation. While the discrete nature and strict bounds of the outcome violate the technical assumptions of the Gaussian family (as the distribution is constrained and not continuous), the approach is justified by its focus on the marginal means effects (AP, IF, OD). The estimator for the regression coefficients derived from this likelihood family is

mathematically equivalent to the Ordinary Least Squares (OLS) estimator under fixed effects and Normal errors (Wooldridge 2010). Crucially, the OLS estimator provides unbiased and consistent estimates of the conditional mean $E(Y \mid X)$ under large samples, even when the residuals are non-Normal or heteroskedastic, relying on asymptotic properties (White 1980). This pragmatic choice allows for directly interpretable coefficients on the original token allocation scale, avoiding the complexity of nonlinear link functions. Furthermore, the inclusion of a rich covariate set and the respondent-level random intercept (u_{i0}) helps to absorb substantial structured variance, mitigating the severity of residual non-Normality that often arises from bounding effects. Respondent-level random intercepts are assumed to follow a normal distribution (Equation 2):

$$u_{i0} \sim \mathcal{N}(0, \sigma^2_{u0}) \tag{2}$$

To capture heterogeneous treatment and cue effects, we include a random intercept and random slopes for T and R at the partisan anchor level. These random effects are assumed to be jointly normal (Equation 3), with intercepts and slopes allowed to be correlated:

$$\begin{pmatrix} v_{a0} \\ v_{aT} \\ v_{aR} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma_a \right), \tag{3}$$

Similarly, we include random intercepts and random slopes for T and R at the country level, also assumed to be jointly normal (Equation 4):

$$\begin{pmatrix} w_{c0} \\ w_{cT} \\ w_{cR} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma_c \right), \tag{4}$$

Both Σ_a and Σ_c are unstructured 3×3 covariance matrices, allowing intercepts and slopes to be estimated as correlated at the anchor and country levels. Residuals are assumed to be normally distributed around 0 (Equation 5):

$$\varepsilon_{riac} \sim \mathcal{N}(0, \sigma^2). \tag{5}$$

The interaction $\beta_3(T_i \times R_{ri})$ is the key parameter: it captures how explicit partisan type ($T_i = 1$) conditions sensitivity to co-, out- and nonpartisan cues (R_{ri}). Because all profile attributes – including the partisan cue – are randomized within respondents, R_{ri} varies as-if at random conditional on A_i, supporting causal interpretation of cue effects within each partisan type. We allow the effect of R_{ri} to vary across partisan anchors (v_{aR}) and across countries (w_{cR}) to capture heterogeneity plausibly structured by party and national context. Put differently, the magnitude of AP and its subcomponent likely varies across parties and across countries, and the random slopes are designed to account for this structure. Similarly, the effect of T_i is allowed to vary across partisan anchors (v_{aT}) and countries (w_{cT}), reflecting that the meaning and salience of explicit attachment differ across parties and settings. Respondent-level heterogeneity in baseline generosity is captured by the random intercept u_{i0}. We do not include respondent-level random slopes for R_{ri} as recommended by Heisig and Schaeffer (2019), because within-respondent variation in partisan relationship categories is insufficient to reliably estimate individual-specific cue sensitivities. Excluding this slope shifts heterogeneity in cue sensitivity to the anchor and country levels, remaining individual-level variation is absorbed by higher-level and residual variances. The correlated random-effects structure enables partial pooling, stabilizing heterogeneous treatment and cue-effect estimates while preserving systematic contextual variation.

We estimate the model in a Bayesian framework via Markov Chain Monte Carlo (MCMC) simulation implemented in Stan (Stan Development Team 2025) using the `brms` package in R (Bürkner 2017, 2018, 2021). This approach propagates uncertainty through the multilevel interaction structure and regularizes the maximal random-effects specification through priors. Our priors are weakly informative and tailored to the 0-10 outcome scale. Fixed-effect parameters (β_k) receive $\mathcal{N}(0, 2)$ priors, which center effects at zero while constraining 95% of prior mass to approximately ± 4 tokens — a substantively large shift (40% of the scale) given the initial findings of Hahm, Hilpert, and König (2024). The model intercept receives a $\mathcal{N}(5, 2)$ prior centered at the midpoint of the token scale. Standard deviations of the random effects σ_{RE} are assigned Half-student-$t(3, 0, 1)$ priors. This distribution is defined exclusively on the positive domain ($[0, \infty)$) and has heav-

ier tails compared to the Half-Normal distribution, allowing for the possibility of larger group-level variability while still shrinking effects towards zero. The residual standard deviation σ is assigned a Half-student-$t(3, 0, 2)$ prior. These priors improve convergence, facilitate shrinkage in a high-dimensional multilevel structure, and reduce the probability of implausible prior-predictive draws (e.g., $Y > 10$).

The randomized conjoint design secures internal validity for profile-level effects (R_{ri} and other Z attributes) within partisan types. The contrast across partisan types (i.e, the effect of R conditional on T) remains observational and is identified under adjustmend for C_i and the multilevel structure. Our hierarchical specification – allowing anchor- and country-varying slopes – and weakly informative priors stabilize these comparisons while transparently reflecting remaining uncertainty in the posterior.

3.4 Quantities of Interest: AP, IF and OD

In this section we define our three substantive quantities of interest — AP and its two constituent components, IF and OD. Our empirical focus lies in comparing these measures across implicit and explicit partisan types. Framed causally, we aim to recover the effect of a hypothetical intervention: holding all covariates constant, how do these quantities of interest change when T is set from 0 (implicit type) to 1 (explicit type)?

We conceptualize AP as the difference in the expected number of tokens allocated to copartisans versus outpartisans in the conjoint games. This follows the approach used in prior work (CHECK Wagner Reiljan), which averages across all political outgroups — a simplification in multi-party systems but one that yields a conservative measure and facilitates cross-national comparison, as both the number of viable parties and the size of the outgroup set vary across countries. Formally, we define affective polarization AP as:

$$AP = E(Y \mid R = Co) - E(Y \mid R = Out) \tag{6}$$

With respect to partisan type, we define our causal estimand as the conditional average treatment effect:

$$\begin{aligned} CATE_{AP} &= AP_{T=1} - AP_{T=0} \\ &= [E(Y \mid R = Co, T = 1) - E(Y \mid R = Out, T = 1)] \\ &\quad - [E(Y \mid R = Co, T = 0) - E(Y \mid R = Out, T = 0)] \end{aligned} \tag{7}$$

To disentangle AP into its components, we exploit the neutral control condition in the conjoint experiment, in which no partisan cue was displayed ($R = \text{None}$). This allows us to estimate the extent to which copartisanship increases token allocations relative to a neutral baseline, conditional on T. We define ingroup favoritism IF as:

$$IF = E(Y \mid R = Co) - E(Y \mid R = None) \tag{8}$$

Refining this with respect to partisan type yields the causal estimand:

$$\begin{aligned} CATE_{IF} &= IF_{T=1} - IF_{T=0} \\ &= [E(Y \mid R = Co, T = 1) - E(Y \mid R = None, T = 1)] \\ &\quad - [E(Y \mid R = Co, T = 0) - E(Y \mid R = None, T = 0)] \end{aligned} \tag{9}$$

The same neutral condition ($R = \text{None}$) also enables estimation of outgroup derogation as the reduction in tokens allocated to outpartisans compared to the baseline. Outgroup derogation OD is defined as:

$$OD = E(Y \mid R = None) - E(Y \mid R = Out) \tag{10}$$

The corresponding causal estimand is:

$$\begin{aligned} CATE_{OD} &= OD_{T=1} - OD_{T=0} \\ &= [E(Y \mid R = None, T = 1) - E(Y \mid R = Out, T = 1)] \\ &\quad - [E(Y \mid R = None, T = 0) - E(Y \mid R = Out, T = 0)] \end{aligned} \tag{11}$$

By focusing on expected values (marginal means), we ensure that our substantive results do not depend on arbitrary choices of reference categories (Leeper, Hobolt, and Tilley 2020). Consistent with our hypotheses stated previously, we interpret all

$CATE_{AP,IF,OD} > 0$ as evidence that respondents with explicit partisan attachments ($T = 1$) display higher AP than implicit types ($T = 0$) (*party-over-policy*). Conversely, we interpret $CATE_{AP,IF,OD} = 0$ as evidence in favor of the *policy-over-party* account, suggesting explicit attachment is not necessary for AP behaviors.

3.5 Covariates and Confounder Adjustment

To block potential backdoor paths between partisan type and allocation behavior (C_i in Figure 2), we adjust for a comprehensive set of individual-level covariates capturing the principal dimensions that may jointly predict explicit partisan attachment and prosocial behavior in experimental settings.

First, we include standard sociodemographic characteristics: age, gender, education, social class, religious affiliation, and urban-rural residence. Second, we control for respondents' political orientations and ideological predispositions, including left-right self-placement, cultural and economic nativism, attitudes toward the EU, and satisfaction with democracy. Third, we account for political interest, engagement, and knowledge. Fourth, we incorporate attitudes toward parties, elites, and democratic processes, capturing broader orientations toward the political system that plausibly shape both group attachment and intergroup conduct. Finally, we include economic evaluations and behavioral despositions (risk-taking and temporal discounting) to adjust for contextual and psychological factors associated with baseline generosity.

Together, these covariates represent substantively plausible confounders and help approximate conditional independence between partisan type and the outcome of interest. Summary statistics for all covariates by partisan type appear in Table 4, and English-language questionnaire items are provided in the appendix.

3.6 Identification Strategy and Causal Assumptions

Our empirical strategy aims to estimate the causal effect of explicit partisan type ($T = 1$) relative to implicit partisan type ($T = 0$) on AP and its subcomponents IF and OD. The

identification approach combines strong internal validity provided by the conjoint experiment for the profile attributes (Z) with an extensive model-based adjustmend strategy for the observed contrast across partisan types (T).

The conjoint design provides strong leverage: all profile attributes (Z) are randomized within each round. Since the respondent-profile relationship (R_{ri}) is a deterministic function of a randomized cue (Z_r^{party}) and the respondent's partisan anchor (A_i), the profile-level variation that drives AP, IF, and OF is as-if random. Consequently, the effect of R (and other randomized attributes) on the token allocation Y_{riac} is causally identified by design, conditional on the respondent's partisan type T.

The key identification challenge is the conditioning of cue effects on partisan type, estimated via the interaction $T \times R$. Identification of this contrast relies on the observational assumption that T is conditionally independent of potential outcomes, given the comprehensive set of covariates C and the full hierarchical structure. Our modeling strategy is explicitly designed to manage this residual confounding.

We formally state the key assumptions and detail our strategy for addressing the principal threat of unobserved confounding:

SUVTA

We assume no interference across respondents and a well-defined treatment. The survey was administered individually and anonymously, rendering cross-respondent spillovers unlikely. Dependence across rounds within respondents is mitigated by randomized profiles and explicitly modeled via the respondent-level random intercepts. Treatment consistency is ensured through a precise coding of T, exclusion of nonpartisan respondents from the analysis, and by allowing its effect to vary across parties and countries.

Conditional Independence (Ignorability) of T

Conditional on the set of observed covariates C and the grouping units (partisan anchor, country), assignment to explicit versus implicit partisan type is assumed intependent of

potential outcomes. While randomization of Z secures identification of the effects of R, the causal interpretation of the Conditional Average Treatment Effects ($CATE$s) across T rests on this ignorability assumption.

Covariate Balance

Observed covariates should be comparably distributed across $T = 1$ and $T = 0$. Descriptive assessments in Table 4 in the appendix indicate a high degree of distributional symmetry across the two groups. The medians are identical for 21 out of 23 scaled political and economic variables, suggesting that the regression adjustment (via $\beta_5 C_i$) starts from a strong, balanced position, substantially mitigating confounding by observed characteristics. Categorical sociodemographic variables also show similar proportions, with the main observed difference being that explicit partisans ($T = 1$) are more likely to be male (53% vs 43%).

Common Support (Overlap)

Both partisan types must occur with positive probability for observed covariate profiles across all grouping units.

The principal threat to the causal interpretation of the $CATE$s is residual confounding: unmeasured factors may jointly influence selection into T and allocation behavior Y. We deploy a multilevel strategy that leverages the repeated-measures design to control for unobserved condounding at all three hierarchical levels:

1. Respondent-level adjustment (u_{i0}): The inclusion of a respondent-level random intercept acts as a powerfull control for all stable, time-invariant, unobserved individual characteristics (e.g., latent personality traits, baseline generosity, or stable intensity of social identity) that may affect both selection into T and the outcome Y. This leverage is fundamental to our observational identification strategy.

2. Partisan anchor-level adjustment (v_{a0}, v_{aT}, v_{aR}): The random effects at the partisan anchor level account for unobserved confounding systematic across party lines.

They absorb average unobserved differences (e.g., party organizational culture or ideological uniformity) that influence both the propensity for explicit attachment (T) and baseline prosocioality (Y). The random slopes for T and R further allow the confounding or effect size to vary by party.

3. Country-level adjustment (w_{c0}, w_{cT}, w_{cR}): The country-level random effects address unobserved confounding operating at the national context level. This accounts for country-specific factors (e.g., general political climate, institutional trust, or overall polarization intensity) that influences both the prevalence of explicit attachment (T) and the general levels of prosocial behavior (Y).

Our hierachical Bayesian model encodes this strategy. All reported estimands are obtained via g-computaion (observed-values approach), averaging posterior expectations over the empirical covariate distribution and relevant grouping units, which propagates uncertainty through the entire multilevel structure.

We are confident in the internal validity of profile-level effects (R and other Z attributes) and of the derived quantities (AP, IF, OD) conditional on T, as these rely on random assignement of Z. The contrast between explicit and implicit partisans — the $CATE$s — is necessarily more assumption-intensive. Our modeling strategy, combining rich covariate adjustment and hierarchical partial pooling, substanially reduces, though cannot eliminate, the possibility of unobserved confounding.

Our goal is to approximate the effect of a hypothetical intervention shifting a respondent's partisan type from implicit ($T = 0$) to explicit ($T = 1$), holding other factors constant. Because T is observational, this interpretation is valid only under the assumptions outlined above. We thus present the resulting estimates as the best model-based causal approximation afforded by the available data and the experimental design.

4 Empirical analysis

We organize the presentation of results in two steps. We begin with pooled analyses based on posterior expected values generated from the fitted hierarchical model. These expected

values are computed for prespecified combinations of respondent's partisan type ($T_i = 0$ implicit, $T_i = 1$ explicit) and partisan relationship category ($R_{ri} \in \{Co, Out, None\}$) in a given game round, averaging over the empirical distribution of covariates and integrating over respondent random effects. From these expected values, we derive the three substantive quantities of interest — affective polarization (AP), ingroup favoritism (IF), and outgroup derogation (OD) — for each value of T. The difference between the quantities under explicit and implicit partisanship yields the $CATE$ for each QoI.

The pooled results sections display, for each QoI, the posterior distributions for explicit and implicit partisans (left panels) alongside the posterior distributions of the corresponding $CATE$ (right panels). All panels visualise full posterior densities along with respective medians, 95% and 99% credible intervals.

We then turn to country-specific analyses, where we recompute expected values separately within each national subsample (again averaging over that country's empirical covariate distribution, but this time incorporating random effects). Because the resulting $CATE$s show little cross-national variation, we additionally report the underlying country-level expected values for each relationship category. These reveal substantial heterogeneity in the levels of AP, even as the implicit-explicit comparison remains consistently close to zero across countries.

All results are presented separately for the dictator and trust games. For scale reference, a one-token difference corresponds to 42.6% of a standard deviation in the dictator game and 40.2% in the trust game.

4.1 The effect of partisan type on affective polarization in Europe

Dictator game

Figure 3 displays the pooled dictator-game estimates. Interpreted within the causal framework introduced above — where the $CATE$ approximates the effect of a hypothetical intervention shifting individuals from *implicit* ($T_i = 0$) to *explicit* ($T_i = 1$) partisan type —

the findings suggest that such an intervention would have, at most, minimal consequences for allocative discrimination.

The overall AP contrast is highly similar across partisan types. Explicit partisans allocate on average 0.99 tokens more to copartisans than to outpartisans (median AP = 0.99; 95%CrI : $0.91, 1.08$), virtually matching the behavior of implicit partisans (median AP = 0.93; 95%CrI : $0.75, 1.10$). The estimated causal effect is therefore small ($CATE_{AP}$ = 0.07; 95%CrI : $-0.11, 0.24$), with roughly three-quarters of its posterior mass concentrated below 0.1 tokens. Although the direction of the posterior mode is consistent with H1a (stronger AP among explicit partisans), the uncertainty dominates, and the key empirical signal is the strength of AP among implicit partisans. Even without self-reported attachment, they display clear and precisely estimated discriminatory allocations — consistent with H1b, the hypothesis that categorization alone suffices to generate behavioral AP.

A similar pattern appears for IF. Explicit partisans show a strong tendency to reward copartisans relative to neutral alters (median IF = 0.54; 95%CrI : $0.46, 0.63$), and implicit partisans exhibit a nearly equivalent tendency (median IF = 0.45; 95%CrI : $0.26, 0.64$). The corresponding causal contrast again remains modest ($CATE_{IF}$ = 0.09; 95%CrI : $-0.10, 0.25$). The posterior probability that explicit partisans would show stronger IF under the hypothetical intervention is 83%, but the implied effect sizes are substantively negligible. Thus, despite the slight directional tendency, the evidence aligns more closely with H2b (no discernible amplification from explicit partisans) than with H2a.

OD yields the clearest indication of equivalence. Explicit partisans withhold 0.45 tokens from outpartisans (95%CrI : $0.36, 0.54$), nearly identical to implicit partisans (\$ median OD = 0.48\$; 95%CrI : $0.35, 0.60$). The causal effect is essentially null ($CATE_{OD} = -0.03$; 95%CrI : $-0.16, 0.10$). If anything, the posterior slightly favors greater derogation among implicit partisans ($Pr(CATE_{OD} < 0) = 0.67$), though again the magnitudes are trivial. These results directly contradict H3a and instead support H3b, which posits no additional derogatory behavior from explicit partisanship.

Viewed jointly, the dictator-game evidence suggests that explicitly endorsing a parti-

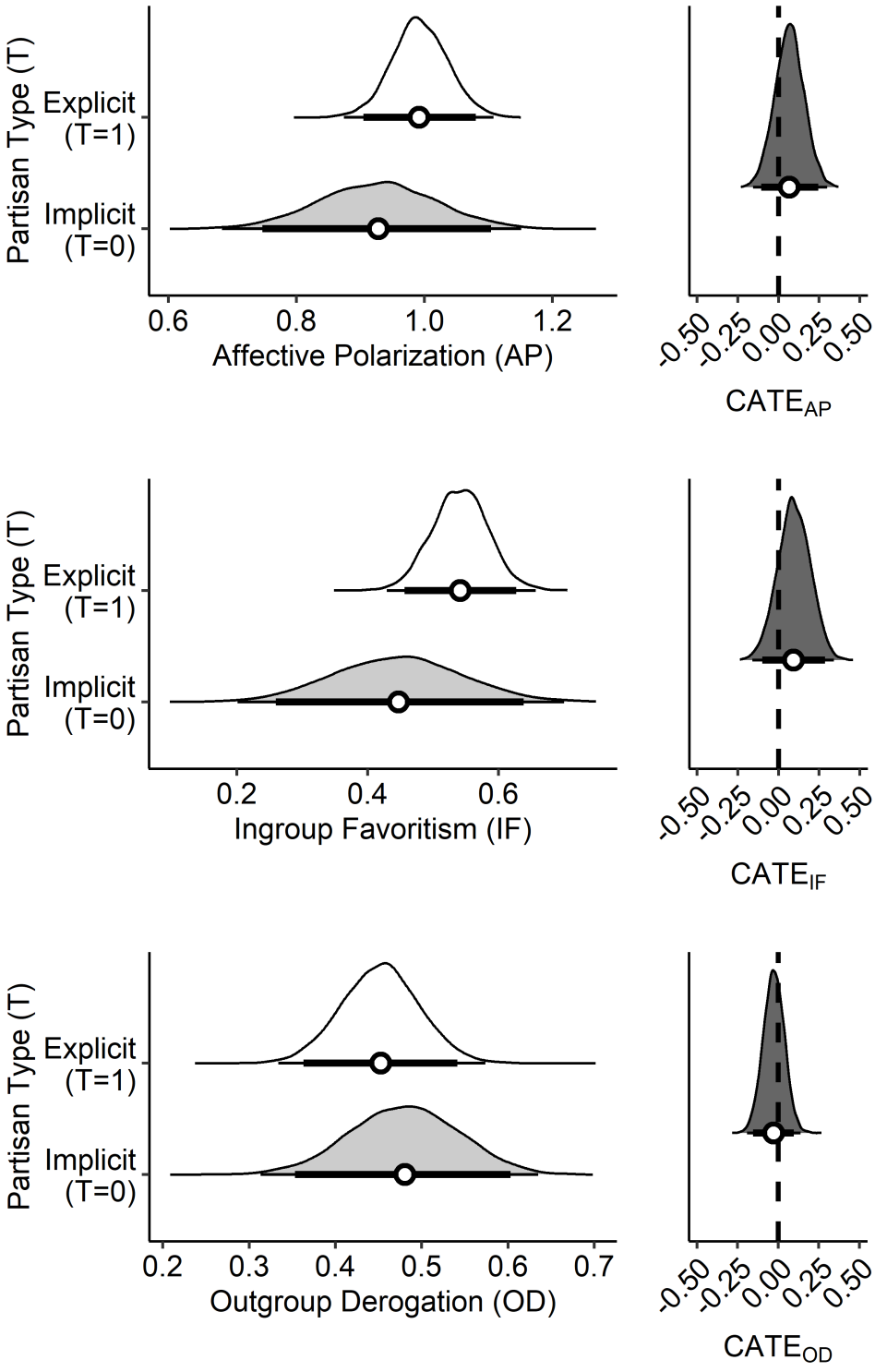

Figure 3: Pooled posterior distributions for AP, IF, and OD and the respective conditional average treatment effects in the dictator game. The figure displays pooled posterior estimates of three derived quantities — affective polarization (AP), ingroup favoritism (IF), and outgroup derogation (OD) — on the token outcome $Y_r iac$ in the dictator game. Rows correspond to AP, IF, and OD. In each row the left panel shows the marginal posterior distirbutions by partisan type (explicit $T_i = 1$, implicit $T_i = 0$), and the right panel shows the corresponding Conditional Average Treatment Effect (CATE), i.e., the difference between explicit and implicit partisans for that quantity. All posterior quantities are obtained from the covariate-adjusted hierarchical model described in Section 3.3 using the observed-values approach (g-computation). The panels depict full posterior densities, points mark posterior medians, thick bars denote 95% credible intervals and thin bars denote 99% credible intervals.

san identity adds little to the behavioral AP dynamics already present among implicit partisans. Posterior medians near zero, symmetric uncertainty, and concentration around substantively small effects all indicate that — under the intervention logic motivating our estimand — explicit identification does not meaningfully amplify partisan discrimination. Instead, the behavioral expression of AP appears grounded in categorization and partisan anchoring, consistent with theoretical accounts that emphasize informational or norm-based mechanisms rather than subjective self-placement. We next examine whether this pattern extends to the trust game, where reciprocal structure may alter the relative contribution of ingroup and outgroup components.

Trust game

Figure 4 summarizes the pooled trust-game results. Although the broader conclusion mirrors the dictator game — limited causal influence of explicit over implicit partisanship — the component patterns differ in informative ways. Relative to the dictator game, the trust game produces (i) no directional signal fo AP, (ii) a reversed pattern for IF, and (iii) the clearest support for the hypothesized effect for OD.

For AP, explicit and implicit partisans behave indistinguishably: both allocate 1.08 tokens more to copartisans (explicit 95%CrI : $1.00, 1.17$; implicit 95%CrI : $0.89, 1.26$). The estimated causal contrast is centered exactly at zero ($CATE_{AP} = 0.00$; 95%CrI : $-0.17, 0.19$), producing a posterior split nearly evenly around the null. This offers no support for H1a in the reciprocal context and is consistent with H1b.

IF, by contrast, exhibits a modest reversal. Explicit partisans reward copartisans by 0.63 tokens (95%CrI : $0.53, 0.72$), whereas implicit partisans display a slightly larger premium (median = 0.75; 95%CrI : $0.53, 0.95$). The resulting causal estimate is negative ($CATE_{IF} = -0.12$; 95%CrI : $-0.32, 0.09$), with the posterior assigning a probability of 0.86 to implicit partisans favoring copartisans *more* strongly. Although small, this direction reverses the dictator-game tendency and suggests that reciprocal incentives can alter the structure of group-based benefits. The evidence therefore contradicts H2a and aligns with H2b.

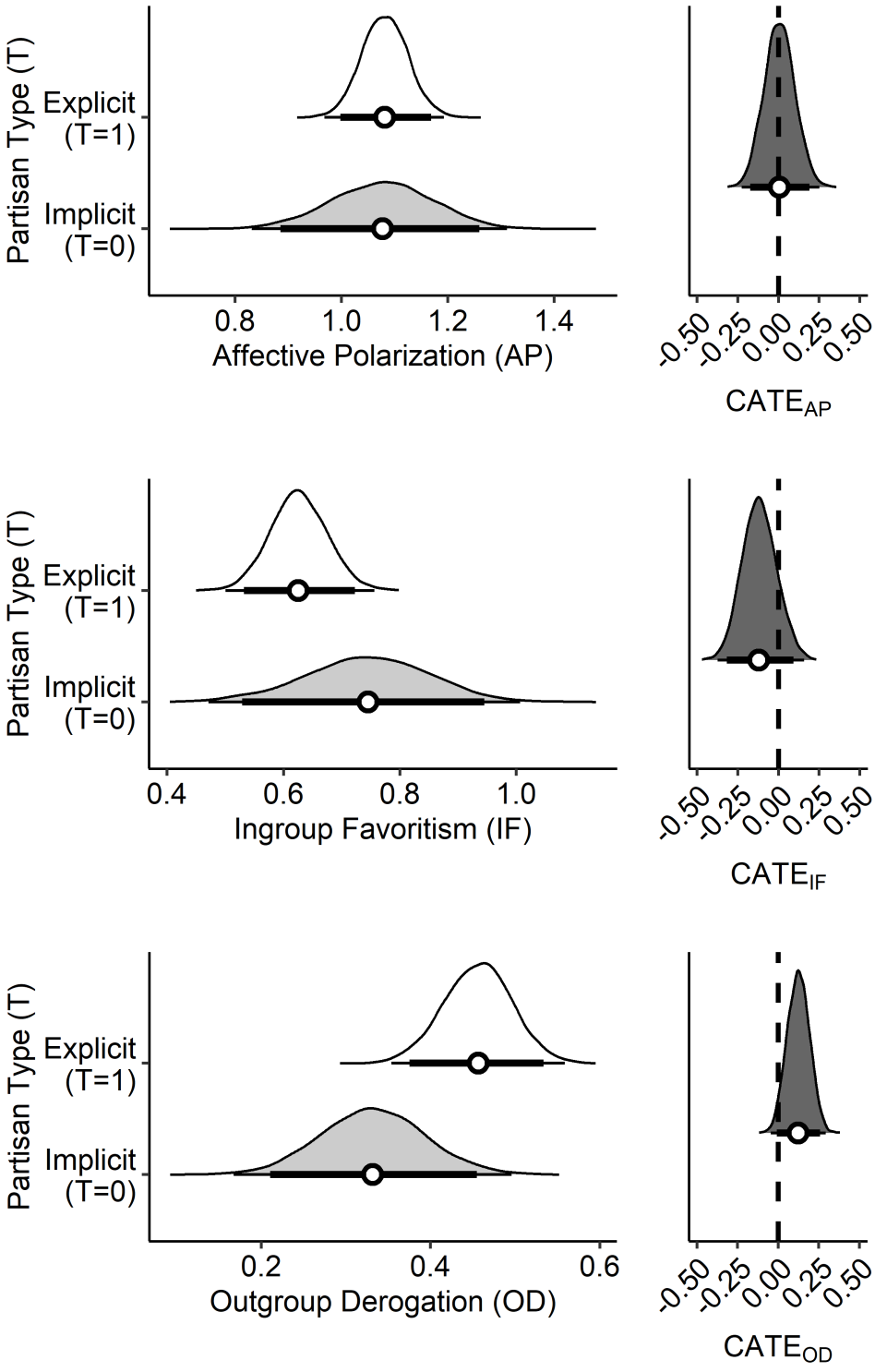

Figure 4: Pooled posterior distributions for AP, IF, and OD and the respective conditional average treatment effects in the trust game. The figure displays pooled posterior estimates of three derived quantities — affective polarization (AP), ingroup favoritism (IF), and outgroup derogation (OD) — on the token outcome $Y_r iac$ in the trust game. Rows correspond to AP, IF, and OD. In each row the left panel shows the marginal posterior distributions by partisan type (explicit $T_i = 1$, implicit $T_i = 0$), and the right panel shows the corresponding Conditional Average Treatment Effects (CATE), i.e., the difference between explicit and implicit partisans for that quantity. All posterior quantities are obtained from the covariate-adjusted hierarchical model described in Section 3.3 using the observed-values approach (g-computation). The curves depict full posterior densities, points mark posterior medians, thick bars denote 95% credible intervals and thin bars denote 99% credible intervals.

Outgroup derogation provides the most supportive trust-game evidence for the hypothesized influence of explicit identification. Explicit partisans withhold 0.46 tokens (95%CrI : $0.38, 0.53$) from outpartisans, compared with 0.33 among implicit partisans (95%CrI : $0.21, 0.46$). The causal contrast is positive ($CATE_{OD} = 0.12$; 95%CrI : $-0.01, 0.26$), and the posterior probability that explicit partisans derogate more is 97%. This is the clearest instance in which results align with H3a. Yet even here, the substantive magnitude remains limited, and implicit partisans still exhibit meaningful OD; the evidence is supportive but not decisive.

Synthesizing across both games, the findings consistently indicate that explicit self-reported partisanship exerts only small, uncertain, and context-dependent causal effects on discriminatory AP behaviors. In several cases the directional tendency aligns with the party-over-policy expectations (H1a–H3a), in others it reverses, and in no instance does the estimated effect approach a substantially large magnitude or produce strong posterior certainty. The lack of convergence across the two behavioral tasks is particularly informative: behaviors attributed to explicit partisan identity do not replicate consistently across allocative and reciprocal environments, undermining the identity-first expectation that subjective attachment reliably intensifies partisan discrimination.

Instead, the pooled findings point to a robust conclusion: individuals who decline to label themselves as partisans nonetheless exhibit levels of AP — including both IF and OD — that closely match those of explicit partisans. Implicit partisanship alone appears sufficient to generate substantial partisan bias in both behavioral contexts under study, as posited by the cue-based and categorization pathways underlying H1b–H3b.

Finally, the trust-game evidence underscores a key measurement implication. Aggregate AP can mask compensating differences in IF and OD. As observed here, similar levels of AP can arise from distinct combinations of its components, potentially obscuring theoretically relevant heterogeneity. Analyses relying solely on copartisan versus outpartisan contrasts may therefore miss meaningful variation in the underlying mechanisms of AP.

4.2 Do explicit and implicit partisans differ across countries?

We next examine whether the effect of partisan type varies across national contexts. If the causal influence of explicit attachment depends on characteristics such as party-system structure, political culture, or the salience of partisan conflict, we would expect meaningfull cross-country variation in the CATE — especially under hypotheses H1a–H3a.

Figure 5 presents the country-specific CATE estimates for the dictator game. Strikingly, each country's posterior distribution is nearly identical to the pooled estimate. No country exhibits a directional deviation, and the uncertainty intervals overlap almost perfectly. The trust-game results in Figure 6 replicate this pattern: the country-level CATEs again mirror the pooled contrasts, with no systematic cross-national dispersion.

Such uniformity initially raises the concern that partial pooling or model specification might have obscured meaningful heterogeneity. To address this, we turn to the conditional expected values (EVs) that underpin the QoIs and their respective CATEs and examine the underlying behavioral patterns directly.

In the dictator game (Figure 7), clear cross-national differences emerge when comparing across rows. Countries vary in baseline giving (as indicated by contributions to "None") and in the magnitude of AP and its components (visible as the relative distances between the shaded posteriors associated with copartisan, neutral, and outpartisan recipients). These differences demonstrate that the model is not surpressing genuine cross-country variation: respondents in some contexts exhibit more AP, IF, or OD than those in others.

However, when comparing across partisan types within each country (the two columns in each panel), the expected values are virtually indistinguishable. In every case, implicit and explicit partisans display the same pattern of allocations, differing mainy in posterior precision (implicit types tend to have wider distributions) rather than in their locations. For example, Ireland exhibits lower overall AP than Spain, but this holds for both implicit and explicit partisans. The same cross-national ordering — and the same within-country equality — appears throughout the sample. This symmetry explains why the CATEs are identical across countries: while countries differ in how much partisans

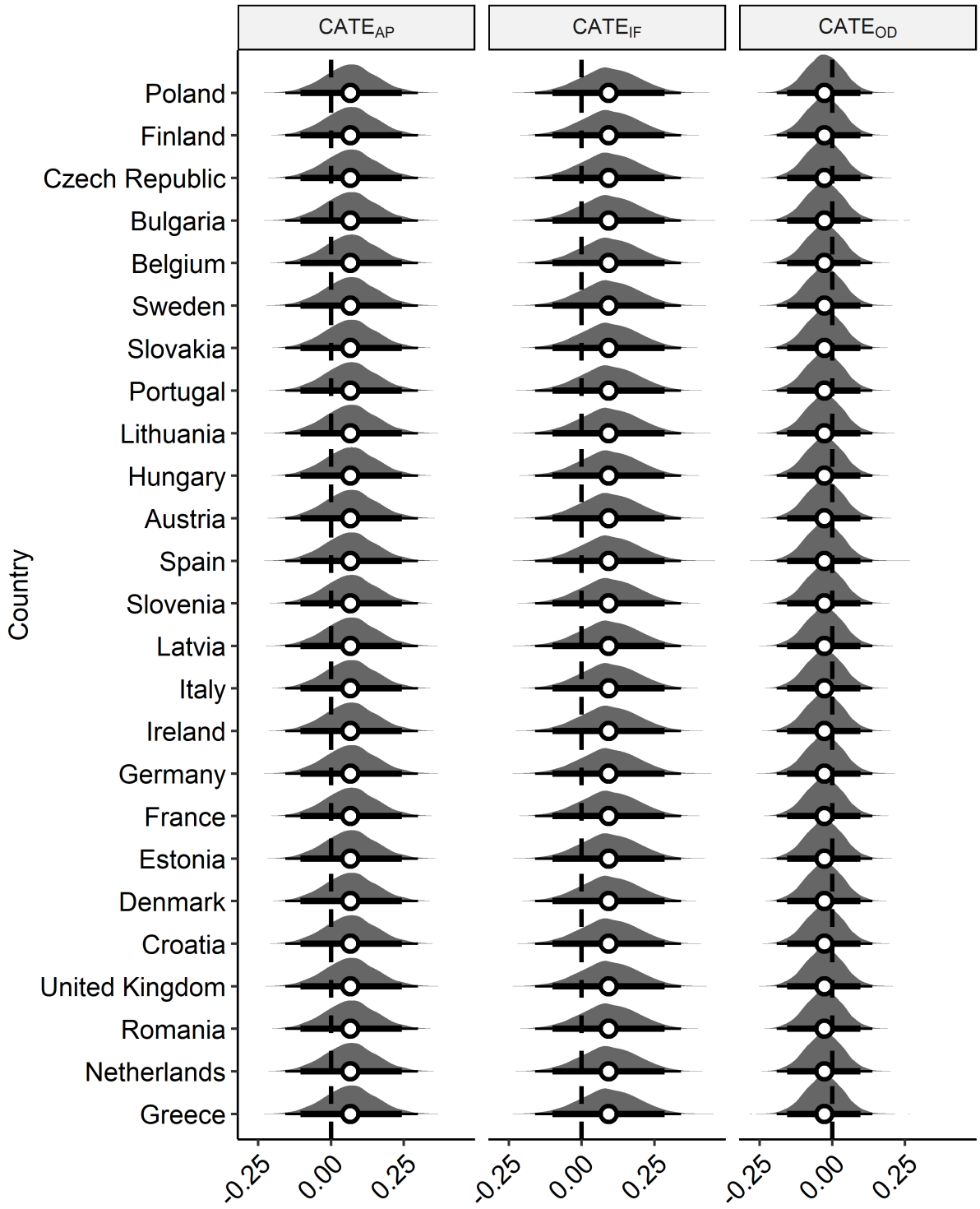

Figure 5: Country-specific CATEs for AP, IF, and OD in the dictator game. The figure displays country-specific posterior distributions of the Conditional Average Treatment Effects (CATEs) of explicit ($T_i = 1$) versus implicit ($T_i = 0$) partisanship on three derived quantities – affective polarization (AP), ingroup favoritism (IF), and outgroup derogation (OD) – measured on the token outcome $Y_r iac$. Columns correspond to $CATE_{AP}$, $CATE_{IF}$, and $CATE_{OD}$, respectively. The y-axis lists countries and the x-axis reports effect sizes in tokens. Positive values indicate larger AP/IF/OD among explicit relative to implicit partisans within a country. All estimates are sampled from the covariate-adjusted hierarchical model (Section 3.3) using the observed-values (g-computation) approach. For each country, we counstruct country-specific datasets that fix T_i and R_{ri} to the focal values, generate posterior predictions including all random effects, and average over the observed distribution of covariates, respondents, and partisan anchors within that country. The curves depict full posterior densities, points mark posterior medians, thick bars denote 95% credible intervals and thin bars denote 99% credible intervals.

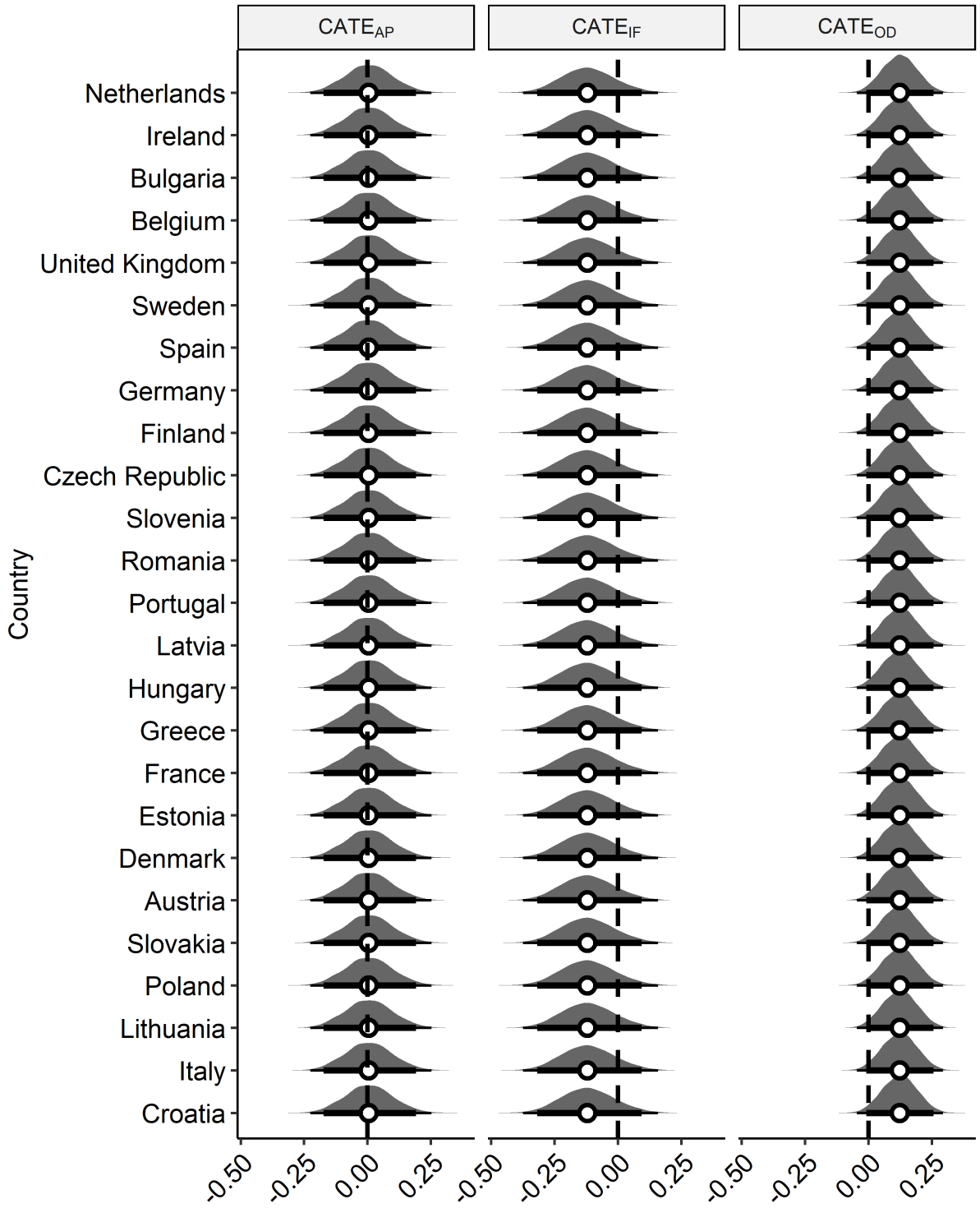

Figure 6: Country-specific CATEs for AP, IF, and OD in the trust game. The figure displays country-specific posterior distributions of the Conditional Average Treatment Effects (CATEs) of explicit ($T_i = 1$) versus implicit ($T_i = 0$) partisanship on three derived quantities – affective polarization (AP), ingroup favoritism (IF), and outgroup derogation (OD) – measured on the token outcome Y_{riac}. Columns correspond to $CATE_{AP}$, $CATE_{IF}$, and $CATE_{OD}$, respectively. The y-axis lists countries and the x-axis reports effect sizes in tokens. Positive values indicate larger AP/IF/OD among explicit relative to implicit partisans within a country. All estimates are sampled from the covariate-adjusted hierarchical model (Section 3.3) using the observed-values (g-computation) approach. For each country, we counstruct country-specific datasets that fix T_i and R_{ri} to the focal values, generate posterior predictions including all random effects, and average over the observed distribution of covariates, respondents, and partisan anchors within that country. The curves depict full posterior densities, points mark posterior medians, thick bars denote 95% credible intervals and thin bars denote 99% credible intervals.

discriminate, they do not differ in how much *additional* discrimination is associated with explicit identification.

The trust-game EVs in Figure 8 reinforce this conclusion. Again, countries differ substantially in aggregate AP and in the composition of IF and OD. Yet within each country, implicit and explicit partisans behave nearly identically, with differences emerging only in the variance of their posterior distributions. The consistency of this pattern across all 25 countries mirrors the pooled estimates.

Taken together, the country-level analyses provide strong evidence that the causal effect of shifting from implicit to explicit partisan type is negligible in every national context examined. Although countries differ in their overall levels of allocative and reciprocal discrimination, the *difference between partisan types* is remarkably stable and effectively zero across all 25 settings. This conclusion is further supported by the hierarchical model: the estimated country-level standard deviation of the partisan-type slope is very small (dictator game: $SD \approx 0.05$; trust game: $SD \approx 0.09$), indicating minimal cross-national dispersion in the implicit–explicit contrast.

In sum, while AP itself varies meaningfully across Europe, the behavioral consequences of explicit versus implicit partisan identification do not. Implicit identification appears sufficient to generate the levels of AP observed in each country, and making partisanship explicit adds little — if anything — to the discriminatory behaviors captured in either experimental task, consistent with the cue-based expectations embodied in H1b–H3b rather than the identity-intensification predictions of H1a–H3a.

5 Robustness

6 Conclusion

This paper asked whether explicit partisan attachment meaningfully conditions affective polarization (AP) in European multiparty systems. Using two behavioral games and a hierarchical reanalysis of a large cross-national conjoint experiment, the evidence points to

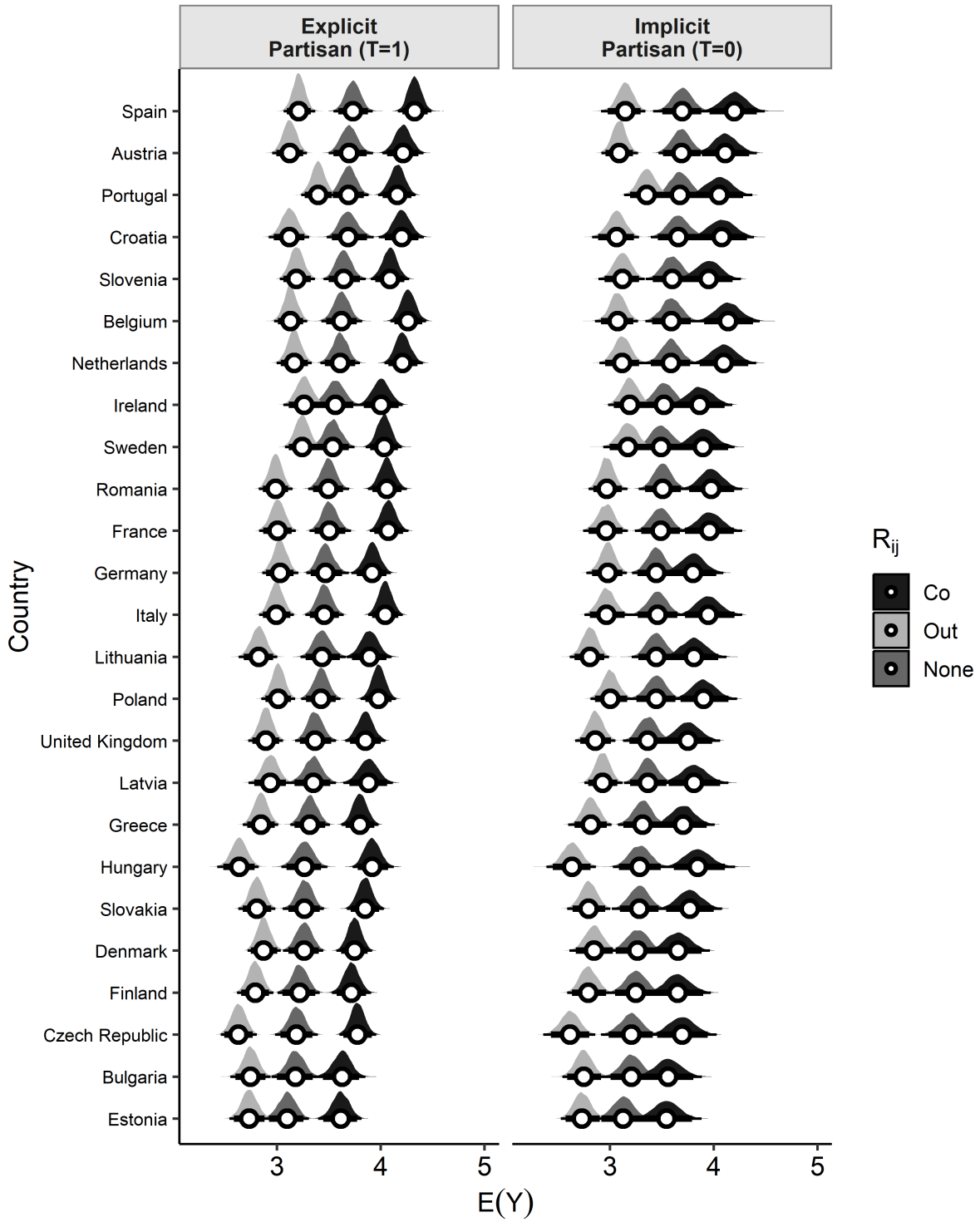

Figure 7: Country-specific posterior expected values by partisan type T_i and relationship R_{ri} in the dictator game. The figure plots country-specific posterior expected token allocations $E(Y_{riac})$ on the 0-10 scale, separatly by partisan type T_i and partisan relationship R_{ri}. The x-axis shows the expected tokens, the y-axis lists countries. Columns split partisan type (left: explicit, $T_i = 1$; right: implicit, $T_i = 0$). Within each column, densities display the posterior distributions for the three relationship conditions $R_{ri} \in Co, Out, None$. Larger separation between Co and Out within a country indicates greater affective polarization (AP). Comparisons across columns reveal how levels vary by partisan type within country. Estimates are obtained from the covariate-adjusted hierarchical model (Section 3.3) using the observed-value (g-computation) approach. For each country, we construct country-specific datasets fixing T_i and R_{ri} to the focal values, generate posterior preditions including all random effects, and average over the observed distribution of covariates, respondents, and partisan anchors within that country. Distributions show full posterior densities, points mark posterior medians, thick and thin horizontal bars denote 95% and 99% credible intervals, respectively.

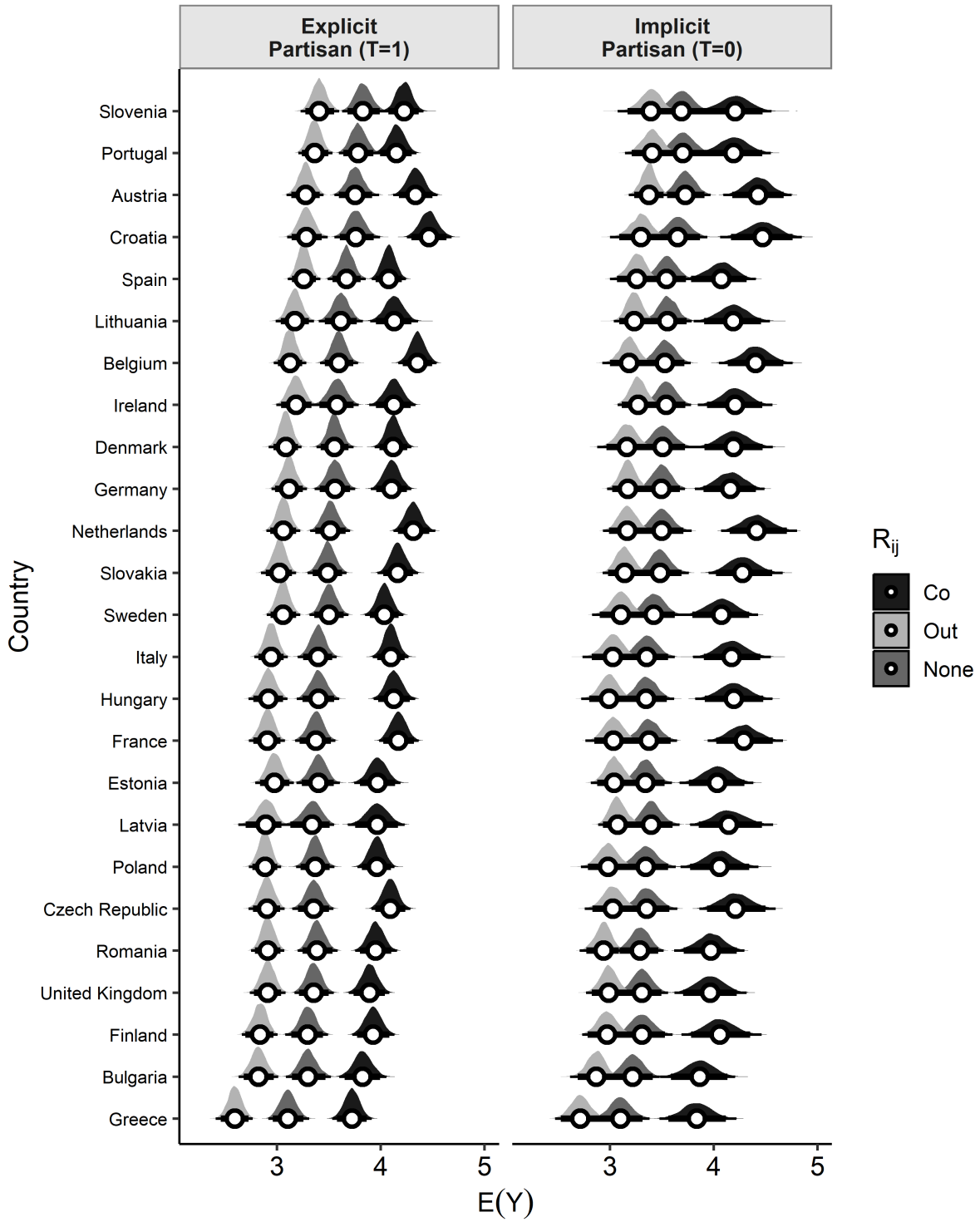

Figure 8: Country-specific posterior expected values by partisan type T_i and relationship R_{ri} in the trust game. The figure plots country-specific posterior expected token allocations $E(Y_{riac})$ on the 0-10 scale, separatly by partisan type T_i and partisan relationship R_{ri}. The x-axis shows the expected tokens, the y-axis lists countries. Columns split partisan type (left: explicit, $T_i = 1$; right: implicit, $T_i = 0$). Within each column, densities display the posterior distributions for the three relationship conditions $R \in Co, Out, None$. Larger separation between Co and Out within a country indicates greater affective polarization (AP). Comparisons across columns reveal how levels vary by partisan type within country. Estimates are obtained from the covariate-adjusted hierarchical model (Section 3.3) using the observed-value (g-computation) approach. For each country, we construct country-specific datasets fixing T_i and R_{ri} to the focal values, generate posterior preditions including all random effects, and average over the observed distribution of covariates, respondents, and partisan anchors within that country. Distributions show full posterior densities, points mark posterior medians, thick and thin horizontal bars denote 95% and 99% credible intervals, respectively.

a consistent answer: across 25 democracies, explicit partisans do not discriminate more strongly than implicit partisans. Conditional average treatment effects are extremely small, posterior mass is tightly concentrated near zero, and the sign of the effect varies across settings and behavioral components. In contrast, implicit partisans — citizens who reject subjective attachment yet reveal a clear partisan anchor — exhibit robust ingroup favoritism (IF) and outgroup derogation (OD) in both games. Behavioral AP thus arises reliably even absent self-reported identity.

These findings carry several implications. First, for measurement and inference, they justify the widespread practice in comparative research of treating vote-anchored respondents and identifiers jointly when estimating behavioral AP. Restricting analyses to explicit partisans risks selecting a shrinking and cross-nationally unstable subgroup, thereby shifting the population of inference in ways that are rarely acknowledged. Distinguishing partisan type helps clarify when such shifts occur and provides a generalizable framework for comparative AP measurement.

Second, the results speak to theories of partisan identity. The standard PID attachment item captures only one dimension of partisanship and appears only weakly related to discriminatory behavior in these tasks. Consistent with recent multidimensional and cue-based theories, behavioral AP in Europe seems to emerge primarily from latent identity, categorization, and substantive inferences that respondents draw from party labels — not from the strength of self-reported attachment. Identity-like behaviors do not presuppose identity-strength, and behavioral forms of AP do not map cleanly onto subjective identification.

Third, the cross-national consistency of the null effect is itself noteworthy. The meaning of explicit identification, and the prevalence of identifiers, vary widely across European electorates, yet the conditional effect of explicit partisanship on AP is remarkably stable: near zero in every country. This pattern suggests a broadly shared mechanism of partisan cue processing that operates similarly across party systems and underscores how multiparty environments structure partisan identity differently than the two-party U.S. context.

Methodologically, the study demonstrates how conjoint experiments combined with Bayesian hierarchical modeling enable researchers to decompose AP into ingroup and outgroup components while accommodating heterogeneous partisan structures. This design yields strong causal leverage on cue effects and provides a template for estimating AP in comparative settings where the meaning of "partisan" varies.

At the same time, limitations remain. Behavioral AP captures only one dimension of a broader construct and it remains to be seen if our findings hold for other indicators, e.g. thermometer ratings, social distance, or elite-focused affect, which may lead to different conclusions. The single-item PID measure is noisy, and richer batteries could sharpen the definition of partisan types. Our design is cross-sectional and cannot adjudicate the dynamics of identity formation. Finally, our interpretation — that the absence of explicit-identity effects supports a policy-over-party rather than party-over-policy mechanism — rests on the assumption that the attachment item accurately captures respondents' self-concepts, an assumption that may not hold uniformly across contexts.

These caveats notwithstanding, the central conclusion is clear. Partisan identity is real, but it is not reducible to reported subjective attachment. Implicit partisanship — the behavioral anchor that shapes categorization and substantive expectations — carries sufficient informational and social content to generate the same discriminatory behaviors as explicit partisans. This challenges identity-only accounts of AP and strengthens the case for multidimensional, cue-driven models of partisan behavior in comparative politics.

7 References

Abramowitz, Alan I. 2010. *The Disappearing Center: Engaged Citizens, Polarization, and American Democracy*. New Haven, CT: Yale University Press.

Abramowitz, Alan I., and Kyle L. Saunders. 2006. "Exploring the Bases of Partisanship in the American Electorate: Social Identity Vs. Ideology." *Political Research Quarterly* 59 (2): 175–87.

———. 2008. "Is Polarization a Myth?" *The Journal of Politics* 70 (2): 542–55. https://doi.org/10.1017/S0022381608080493.

Adams, James, David Bracken, Noam Gidron, Will Horne, Diana Z O'Brien, and Kaitlin Senk. 2023. "Can't We All Just Get Along? How Women MPs Can Ameliorate Affective Polarization in Western Publics." *American Political Science Review* 117 (1): 318–24. https://doi.org/10.1017/S0003055422000491.

Allport, Gordon W. 1954. *The Nature of Prejudice*. Reading, MA: Addison-Wesley.

Areal, J., and E. Harteveld. 2024. "Vertical Vs Horizontal Affective Polarization: Disentangling Feelings Towards Elites and Voters." *Electoral Studies* 90. https://doi.org/10.1016/j.electstud.2024.102814.

Bail, Christopher A., Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. "Exposure to Opposing Views on Social Media Can Increase Political Polarization." *Proceedings of the National Academy of Sciences* 115 (37): 9216–21. https://doi.org/10.1073/pnas.1804840115.

Bakker, Bert N, and Yphtach Lelkes. 2024. "Putting the Affect into Affective Polarisation." *Cognition and Emotion* 38 (4): 418–36. https://doi.org/10.1080/02699931.2024.2362366.

Bansak, Kirk, Jens Hainmueller, Daniel J. Hopkins, Teppei Yamamoto, James N. Druckman, and Donald P. Green. 2021. "Conjoint Survey Experiments." In *Advances in Experimental Political Science*, edited by James N. Druckman and Donald P. Green, 19:19–41. Cambridge University Press Cambridge. https://doi.org/10.1017/9781108777919.004.

Bantel, Ivo. 2023. "Camps, Not Just Parties. The Dynamic Foundations of Affective Polarization in Multi-Party Systems." *Electoral Studies* 83: 102614. https://doi.org/10.1016/j.electstud.2023.102614.

Berntzen, Lars Eric, H. Kelsall, and E. Harteveld. 2023. "Consequences of Affective Polarization: Avoidance, Intolerance and Support for Violence in the United Kingdom and Norway." *European Journal of Political Research*. https://doi.org/10.1111/1475-6765.12623.

Berntzen, Lars Erik. 2025. "Affective Polarization and Political Violence." In *Handbook of Affective Polarization*, edited by Mariano Torcal and Eelco Harteveld, 414–28. Edward Elgar Publishing. https://doi.org/10.4337/9781035310609.00043.

Bougher, Lori D. 2017. "The Correlates of Discord." *Political Behavior* 39 (3): 731–62. https://doi.org/10.1007/s11109-016-9377-1.

Boxell, Levi, Matthew Gentzkow, and Jesse M. Shapiro. 2024. "Cross-Country Trends in Affective Polarization." *Review of Economics and Statistics* 106 (2): 557–65. https://doi.org/10.1162/rest_a_01160.

Brewer, Marilynn B. 1999. "The Psychology of Prejudice: Ingroup Love and Outgroup Hate?" *Journal of Social Issues* 55 (3): 429–44. https://doi.org/10.1111/0022-4537.00126.

Broockman, D. E., J. L. Kalla, and S. J. Westwood. 2023. "Does Affective Polarization Undermine Democratic Norms or Accountability? Maybe Not." *American Journal of Political Science* 67 (3): 1–21. https://doi.org/10.1111/ajps.12719.

Bürkner, Paul-Christian. 2017. "brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software* 80 (1): 1–28. https://doi.org/10.18637/jss.v080.i01.

———. 2018. "Advanced Bayesian Multilevel Modeling with the R Package brms." *The R Journal* 10 (1): 395–411. https://doi.org/10.32614/RJ-2018-017.

———. 2021. "Bayesian Item Response Modeling in R with brms and Stan." *Journal of Statistical Software* 100 (5): 1–54. https://doi.org/10.18637/jss.v100.i05.

Campos, Nicolas, and Christopher Federico. 2025. "A New Measure of Affective Polarization." *American Political Science Review*, 1–19. https://doi.org/10.1017/S0003055425000255.

Carlin, Ryan E., and Gregory J. Love. 2025. "Measuring Affective Polarization." In *Handbook of Affective Polarization*, edited by Mariano Torcal and Eelco Harteveld, 52–68. Edward Elgar Publishing. https://doi.org/10.4337/9781035310609.00010.

Dafoe, Allan, Baobao Zhang, and Devin Caughey. 2018. "Information Equivalence in Survey Experiments." *Political Analysis* 26 (4): 399–416. https://doi.org/10.1017/pan.2018.9.

Dias, Nicholas C, Yphtach Lelkes, and Jacob Pearl. 2025. "American Partisans Vastly Under-Estimate the Diversity of Other Partisans' Policy Attitudes." *Political Science Research and Methods* 13 (3): 725–35. https://doi.org/10.1017/psrm.2024.36.

Dias, Nicholas, and Yphtach Lelkes. 2022. "The Nature of Affective Polarization: Disentangling Policy Disagreement from Partisan Identity." *American Journal of Political Science* 66 (3): 775–90. https://doi.org/10.1111/ajps.12628.

Diermeier, Daniel, and Christopher Li. 2019. "Partisan Affect and Elite Polarization." *American Political Science Review* 113 (1): 277–81. https://doi.org/10.1017/S0003055418000655.

Druckman, James N, Donald P Green, and Shanto Iyengar. 2023. "Does Affective Polarization Contribute to Democratic Backsliding in America?" *The ANNALS of the American Academy of Political and Social Science* 708 (1): 137–63. https://doi.org/10.1177/00027162241228952.

Druckman, James, and Matthew Levendusky. 2019. "What Do We Measure When We Measure Affective Polarization?" *Public Opinion Quarterly* 83 (1): 114–22. https://doi.org/10.1093/poq/nfz003.

Elliott, Kevin J. 2024. "What Is It Like to Be a Partisan? Measures of Partisanship and Its Value for Democracy." *Perspectives on Politics* 22 (3): 584–98. https://doi.org/10.1017/S153759272300289X.

Ferreira da Silva, Frederico, and Diego Garzia. 2024. "Affective Polarization Towards Parties and Leaders, and Electoral Participation in 13 Parliamentary Democracies, 1980–2019." *Public Opinion Quarterly* 88 (4): 1234–48. https://doi.org/10.1093/poq/nfae053.

Finkel, Eli J, Christopher A Bail, Mina Cikara, Peter H Ditto, Shanto Iyengar, Samara Klar, Lilliana Mason, et al. 2020. "Political Sectarianism in America." *Science* 370 (6516): 533–36. https://doi.org/10.1126/science.abe1715.

Fiorina, Morris P., Samuel A. Abrams, and Jeremy C. Pope. 2008. "Polarization in the American Public: Misconceptions and Misreadings." *The Journal of Politics* 70 (2): 556–60.

Fiorina, Morris P., and Samuel J. Abrams. 2008. "Political Polarization in the American Public." *Annual Review of Political Science* 11: 563–88.

Fiorina, Morris P., Samuel J. Abrams, and Jeremy C. Pope. 2006. *Culture War? The Myth of a Polarized America*. 2nd ed. Pearson Longman.

Fowler, Anthony, Seth J Hill, Jeffrey B Lewis, Chris Tausanovitch, Lynn Vavreck, and Christopher Warshaw. 2023. "Moderates." *American Political Science Review* 117 (2): 643–60. https://doi.org/10.1017/S0003055422000818.

Garzia, Diego, Frederico Ferreira da Silva, and Simon Maye. 2023. "Affective Polarization in Comparative and Longitudinal Perspective." *Public Opinion Quarterly* 87 (1): 219–31. https://doi.org/10.1093/poq/nfad004.

Gidrom, Noam, James Adams, and Will Horne. 2020. *American Affective Polarization in Comparative Perspective*. Cambridge: University of Cambridge Press. https://doi.org/10.1017/9781108914123.

Graham, Matthew H., and Milan W. Svolik. 2020. "Democracy in America? Partisanship, Polarization, and the Robustness of Support for Democracy in the United States." *American Political Science Review* 114 (2): 392–409. https://doi.org/10.1017/S0003055420000052.

Greene, Steven. 1999. "Understanding Party Identification." *Political Psychology* 20 (2):

393–403. https://doi.org/10.1111/0162-895X.00150.

Hahm, Hyeonho, David Hilpert, and Thomas König. 2024. "Divided We Unite: The Nature of Partyism and the Role of Coalition Partnership in Europe." *American Political Science Review* 118 (1): 69–87. https://doi.org/10.1017/S0003055423000266.

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. https://doi.org/10.1093/pan/mpt024.

Hawkins, Carlee Beth, and Brian A Nosek. 2012. "Motivated Independence? Implicit Party Identity Predicts Political Judgments Among Self-Proclaimed Independents." *Personality and Social Psychology Bulletin* 38 (11): 1437–52. https://doi.org/10.1177/0146167212452313.

Heisig, Jan Paul, and Merlin Schaeffer. 2019. "Why You Should Always Include a Random Slope for the Lower-Level Variable Involved in a Cross-Level Interaction." *European Sociological Review* 35 (2): 258–79. https://doi.org/10.1093/esr/jcy053.

Hobolt, Sara B, Katharina Lawall, and James Tilley. 2024. "The Polarizing Effect of Partisan Echo Chambers." *American Political Science Review* 118 (3): 1464–79. https://doi.org/10.1017/S0003055423001211.

Horiuchi, Yusaku, Zachary Markovich, and Teppei Yamamoto. 2022. "Does Conjoint Analysis Mitigate Social Desirability Bias?" *Political Analysis* 30 (4): 535–49. https://doi.org/10.1017/pan.2021.30.

Huber, Gregory A., and Neil Malhotra. 2017. "Political Homophily in Social Relationships: Evidence from Online Dating Behavior." *The Journal of Politics* 79 (1): 269–83. https://doi.org/10.1086/687533.

Huddy, Leonie. 2013. "Group Identity and Political Cohesion." In *The Oxford Handbook of Political Psychology*, edited by Leonie Huddy, David O. Sears, and Jack S. Levy. Oxford University Press.

Huddy, Leonie, and Alexa Bankert. 2017. "Political Partisanship as a Social Identity." In *Oxford Research Encyclopedia of Politics*. https://doi.org/10.1093/acrefore/9780190228637.013.250.

Huddy, Leonie, Alexa Bankert, and Caitlin Davies. 2018. "Expressive Versus Instrumental Partisanship in Multiparty European Systems." *Political Psychology* 39: 173–99. https://doi.org/10.1111/pops.12482.

Huddy, Leonie, Lilliana Mason, and Lene Aarøe. 2015. "Expressive Partisanship: Campaign Involvement, Political Emotion, and Partisan Identity." *American Political Science Review* 109 (1): 1–17. https://doi.org/10.1017/S0003055414000604.

Iyengar, Shanto, and Masha Krupenkin. 2018. "The Strengthening of Partisan Affect." *Political Psychology* 39: 201–18. https://doi.org/10.1111/pops.12487.

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood. 2019. "The Origins and Consequences of Affective Polarization in the United States." *Annual Review of Political Science* 22 (1): 129–46. https://doi.org/10.1146/annurev-polisci-051117-073034.

Iyengar, Shanto, Gaurav Sood, and Yphtach Lelkes. 2012. "Affect, Not Ideology: A Social Identity Perspective on Polarization." *Public Opinion Quarterly* 76 (3): 405–31. https://doi.org/10.1093/poq/nfs038.

Iyengar, Shanto, and Markus Wagner. 2025. "Conceptualizing Affective Polarization." In *Handbook of Affective Polarization*, edited by Mariano Torcal and Eelco Harteveld, 22–34. Edward Elgar Publishing. https://doi.org/10.4337/9781035310609.00007.

Iyengar, Shanto, and Sean J Westwood. 2015. "Fear and Loathing Across Party Lines: New Evidence on Group Polarization." *American Journal of Political Science* 59 (3): 690–707. https://doi.org/10.1111/ajps.12152.

Jenke, Libby. 2024. "Affective Polarization and Misinformation Belief." *Political Behavior* 46 (2): 825–84. https://doi.org/10.1007/s11109-022-09851-w.

Kalla, Joshua L, and David E Broockman. 2022. "Voter Outreach Campaigns Can Reduce Affective Polarization Among Implementing Political Activists: Evidence from Inside Three Campaigns." *American Political Science Review* 116 (4): 1516–22. https://doi.org/10.1017/S0003055422000132.

Kalmoe, Nathan P., and Lilliana Mason. 2022a. "A Holistic View of Conditional American Support for Political Violence." *Proceedings of the National Academy of Sciences* 119 (32): e2207237119. https://doi.org/10.1073/pnas.2207237119.

———. 2022b. *Radical American Partisanship: Mapping Violent Hostility, Its Causes, and the Consequences for Democracy*. Chicago, IL: University of Chicago Press.

Kekkonen, Arto, Aleksi Suuronen, Daniel Kawecki, and Kim Strandberg. 2022. "Puzzles in Affective Polarization Research: Party Attitudes, Partisan Social Distance, and Multiple Party Identification." *Frontiers in Political Science* 4: 920567. https://doi.org/10.3389/fpos.2022.920567.

Kekkonen, Arto, and Tuomas Ylä-Anttila. 2021. "Affective Blocs: Understanding Affective Polarization in Multiparty Systems." *Electoral Studies* 72: 102367. https://doi.org/10.1016/j.electstud.2021.102367.

Kinder, Donald R., and Nathan P. Kalmoe. 2017. *Neither Liberal nor Conservative: Ideological Innocence in the American Public*. Chicago, IL: University of Chicago Press.

Kingzette, Jon, James N. Druckman, Samara Klar, Yanna Krupnikov, Matthew Levendusky, and John Barry Ryan. 2021. "How Affective Polarization Undermines Support for Democratic Norms." *Public Opinion Quarterly* 85 (2): 663–77. https://doi.org/10.1093/poq/nfab029.

Klar, Samara, and Yanna Krupnikov. 2016. *Independent Politics*. Cambridge: University of Cambridge Press.

Klar, Samara, Yanna Krupnikov, and John Barry Ryan. 2018. "Affective Polarization or Partisan Disdain? Untangling a Dislike for the Opposing Party from a Dislike of Partisanship." *Public Opinion Quarterly* 82 (2): 379–90. https://doi.org/10.1093/poq/nfy014.

Lee, Amber Hye-Yon. 2022. "Social Trust in Polarized Times: How Perceptions of Political Polarization Affect Americans' Trust in Each Other." *Political Behavior* 44 (3): 1533–54. https://doi.org/10.1007/s11109-022-09787-1.

Lee, Sangwon, Jihyang Choi, and Chloe Ahn. 2025. "Hate Prompts Participation: Examining the Dynamic Relationship Between Affective Polarization and Political Participation." *New Media & Society* 27 (1): 443–61. https://doi.org/10.1177/14614448231177301.

Leeper, Thomas J, Sara B Hobolt, and James Tilley. 2020. "Measuring Subgroup Preferences in Conjoint Experiments." *Political Analysis* 28 (2): 207–21. https://doi.org/10.1017/pan.2019.30.

Levendusky, Matthew S. 2018. "Americans, Not Partisans: Can Priming American National Identity Reduce Affective Polarization?" *The Journal of Politics* 80 (1): 59–70. https://doi.org/10.1086/693987.

Levendusky, Matthew S. 2013. "Why Do Partisan Media Polarize Viewers?" *American Journal of Political Science* 57 (3): 611–23. https://doi.org/10.1111/ajps.12008.

Luttig, Matthew D. 2017. "Authoritarianism and Affective Polarization: A New View on the Origins of Partisan Extremism." *Public Opinion Quarterly* 81 (4): 866–95. https://doi.org/10.1093/poq/nfx023.

———. 2018. "The 'Prejudiced Personality' and the Origins of Partisan Strength, Af-

fective Polarization, and Partisan Sorting." *Political Psychology* 39: 239–56. https://doi.org/10.1111/pops.12484.

Mason, Lilliana. 2013. "The Rise of Uncivil Agreement: Issue Versus Behavioral Polarization in the American Electorate." *American Behavioral Scientist* 57 (1): 140–59. https://doi.org/10.1177/0002764212463363.

———. 2015. "'I Disrespectfully Agree': The Differential Effects of Partisan Sorting on Social and Issue Polarization." *American Journal of Political Science* 59 (1): 128–45. https://doi.org/10.1111/ajps.12089.

———. 2016. "A Cross-Cutting Calm: How Social Sorting Drives Affective Polarization." *Public Opinion Quarterly* 80 (S1): 351–77. https://doi.org/10.1093/poq/nfw001.

———. 2018a. "Ideologues Without Issues: The Polarizing Consequences of Ideological Identities." *Public Opinion Quarterly* 82 (S1): 866–87. https://doi.org/10.1093/poq/nfy005.

———. 2018b. *Uncivil Agreement: How Politics Became Our Identity*. The University of Chicago Press.

Mason, Lilliana, and Julie Wronski. 2018. "One Tribe to Bind Them All: How Our Social Group Attachments Strengthen Partisanship." *Political Psychology* 39: 257–77. https://doi.org/10.1111/pops.12485.

McCarty, Nolan, Keith T. Poole, and Howard Rosenthal. 2006. *Polarized America: The Dance of Ideology and Unequal Riches*. Cambridge, MA: MIT Press.

McConnell, Christopher, Yotam Margalit, Neil Malhotra, and Matthew Levendusky. 2018. "The Economic Consequences of Partisanship in a Polarized Era." *American Journal of Political Science* 62 (1): 5–18. https://doi.org/10.1111/ajps.12330.

Mernyk, Joseph S, Sophia L Pink, James N Druckman, and Robb Willer. 2022. "Correcting Inaccurate Metaperceptions Reduces Americans' Support for Partisan Violence." *Proceedings of the National Academy of Sciences* 119 (16): e2116851119. https://doi.org/10.1073/pnas.2116851119.

Moore-Berg, Samantha L, Lee-Or Ankori-Karlinsky, Boaz Hameiri, and Emile Bruneau. 2020. "Exaggerated Meta-Perceptions Predict Intergroup Hostility Between American Political Partisans." *Proceedings of the National Academy of Sciences* 117 (26): 14864–72. https://doi.org/10.1073/pnas.2001263117.

Orr, Lilla V, Anthony Fowler, and Gregory A Huber. 2023. "Is Affective Polarization Driven by Identity, Loyalty, or Substance?" *American Journal of Political Science* 67 (4): 948–62. https://doi.org/10.1111/ajps.12796.

Orr, Lilla V, and Gregory A Huber. 2020. "The Policy Basis of Measured Partisan Animosity in the United States." *American Journal of Political Science* 64 (3): 569–86. https://doi.org/10.1111/ajps.12498.

Petrocik, John Richard. 2009. "Measuring Party Support: Leaners Are Not Independents." *Electoral Studies* 28 (4): 562–72. https://doi.org/10.1016/j.electstud.2009.05.022.

Poole, Keith T, and Howard Rosenthal. 1984. "The Polarization of American Politics." *The Journal of Politics* 46 (4): 1061–79. https://doi.org/10.2307/2131242.

Reiljan, Andres. 2020. "'Fear and Loathing Across Party Lines'(also) in Europe: Affective Polarisation in European Party Systems." *European Journal of Political Research* 59 (2): 376–96. https://doi.org/10.1111/1475-6765.12351.

Reiljan, Andres, Diego Garzia, Frederico Ferreira Da Silva, and Alexander H Trechsel. 2024. "Patterns of Affective Polarization Toward Parties and Leaders Across the Democratic World." *American Political Science Review* 118 (2): 654–70. https://doi.org/10.1017/S0003055423000485.

Reiljan, Andres, and Alexander Ryan. 2021. "Ideological Tripolarization, Partisan Tribal-

ism and Institutional Trust: The Foundations of Affective Polarization in the Swedish Multiparty System." *Scandinavian Political Studies* 44 (2): 195–219. https://doi.org/10.1111/1467-9477.12194.

Roccas, Sonia, and Marilynn B. Brewer. 2002. "Social Identity Complexity." *Personality and Social Psychology Review* 6 (2): 88–106. https://doi.org/10.1207/S15327957PSPR0602_01.

Rogowski, Jon C., and Joseph L. Sutherland. 2016. "How Ideology Fuels Affective Polarization." *Political Behavior* 38: 485–508. https://doi.org/10.1007/s11109-015-9323-7.

Röllicke, L. 2023. "Polarisation, Identity and Affect - Conceptualising Affective Polarisation in Multi-Party Systems." *Electoral Studies* 85. https://doi.org/10.1016/j.electstud.2023.102655.

Rothers, Adrian. 2025. "The Poles in Polarization: Social Categorization and Affective Polarization in Multiparty Systems." *Electoral Studies* 95: 102908. https://doi.org/10.1016/j.electstud.2025.102908.

Sherif, Muzafer, O J Harvey, B Jack White, William R Hood, and Carolyn W Sherif. 1961. *Intergroup Conflict and Cooperation: The Robbers Cave Experiment*. Norman, OK: University Book Exchange.

Stan Development Team. 2025. "RStan: The R Interface to Stan." https://mc-stan.org/.

Tajfel, Henri. 1974. "Social Identity and Intergroup Behaviour." *Social Science Information* 13 (2): 65–93.

Tajfel, Henri, Michael G. Billig, Robert P. Bundy, and Claude Flament. 1971. "Social Categorization and Intergroup Behaviour." *European Journal of Social Psychology* 1 (2): 149–78.

Tajfel, Henri, and John Turner. 2004. "The Social Identity Theory of Intergroup Behavior." In *Political Psychology: Key Readings*, edited by John Jost and Jim Sidanius, 276–93. New York: Psychology Press.

Tajfel, H., and J. Turner. 1979. "An Integrative Theory of Intergroup Conflict." In *The Social Psychology of Intergroup Relation*, edited by William G. Austin and Stephen Worchel, 33–47. Brooks/Cole.

Theodoridis, Alexander G. 2017. "Me, Myself, and (i),(d), or (r)? Partisanship and Political Cognition Through the Lens of Implicit Identity." *The Journal of Politics* 79 (4): 1253–67. https://doi.org/10.1086/692738.

Torcal, Mariano, and Josep Maria Comellas. 2025. "Operationalizing Affective Polarization in Multiparty Systems." In *Handbook of Affective Polarization*, edited by Mariano Torcal and Eelco Harteveld, 69–87. Edward Elgar Publishing. https://doi.org/10.4337/9781035310609.00011.

Torcal, Mariano, Andres Reiljan, and Lisa Zanotti. 2023. "Affective Polarization in Comparative Perspective." *Frontiers in Political Science* 5: 1112238. https://doi.org/10.3389/fpos.2023.1112238.

Törnberg, Petter. 2022. "How Digital Media Drive Affective Polarization Through Partisan Sorting." *Proceedings of the National Academy of Sciences* 119 (42): e2207159119. https://doi.org/10.1073/pnas.2207159119.

Voelkel, Jan G, James Chu, Michael N Stagnaro, Joseph S Mernyk, Chrystal Redekopp, Sophia L Pink, James N Druckman, David G Rand, and Robb Willer. 2023. "Interventions Reducing Affective Polarization Do Not Necessarily Improve Anti-Democratic Attitudes." *Nature Human Behaviour* 7 (1): 55–64. https://doi.org/10.1038/s41562-022-01466-9.

Wagner, Markus. 2021. "Affective Polarization in Multiparty Systems." *Electoral Studies* 69: 102199. https://doi.org/10.1016/j.electstud.2020.102199.

Webster, Steven W., and Alan I. Abramowitz. 2017. "The Ideological Foundations of

Affective Polarization in the US Electorate." *American Politics Research* 45 (4): 621–47. https://doi.org/10.1177/1532673X17703.

West, Emily A, and Shanto Iyengar. 2022. "Partisanship as a Social Identity." *Political Behavior* 44: 807–38. https://doi.org/10.1007/s11109-020-09637-y.

Westwood, Sean J, Justin Grimmer, Matthew Tyler, and Clayton Nall. 2022a. "Current Research Overstates American Support for Political Violence." *Proceedings of the National Academy of Sciences* 119 (12): e2116870119. https://doi.org/10.1073/pnas.2116870119.

———. 2022b. "Reply to Kalmoe and Mason: The Pitfalls of Using Surveys to Measure Low-Prevalence Attitudes and Behavior." *Proceedings of the National Academy of Sciences* 119 (32): e2207584119. https://doi.org/10.1073/pnas.2207584119.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica*, 817–38. https://doi.org/10.2307/1912934.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. MIT press.

8 Appendix

8.1 Sample descriptives

Table 1

Country	N	Percent
Austria	1277	0.04
Belgium	1305	0.04
Bulgaria	982	0.03
Croatia	1240	0.04
Czech Republic	1135	0.04
Denmark	1200	0.04
Estonia	944	0.03
Finland	1160	0.04
France	1156	0.04
Germany	1188	0.04
Greece	1161	0.04
Hungary	986	0.03
Ireland	1061	0.04
Italy	1172	0.04
Latvia	1148	0.04
Lithuania	1265	0.04
Netherlands	1221	0.04
Poland	1198	0.04
Portugal	1187	0.04
Romania	1480	0.05
Slovakia	1297	0.04
Slovenia	1135	0.04
Spain	1396	0.05
Sweden	1254	0.04

Table 1

Country	N	Percent
United Kingdom	1279	0.04
Total	29827	0.99

Table 2

meta_country	1	2	3
Austria	645	630	2
Belgium	729	574	2
Bulgaria	463	518	1
Croatia	545	694	1
Czech Republic	515	618	2
Denmark	690	508	2
Estonia	331	611	2
Finland	586	567	7
France	562	594	0
Germany	587	597	4
Greece	588	572	1
Hungary	493	492	1
Ireland	481	577	3
Italy	603	569	0
Latvia	415	733	0
Lithuania	462	803	0
Netherlands	642	577	2
Poland	540	658	0
Portugal	593	593	1
Romania	837	641	2
Slovakia	550	746	1
Slovenia	569	566	0
Spain	677	718	1
Sweden	648	602	4
United Kingdom	635	641	3
Total	14386	15399	42

8.2 Experimental setup

Before the behavioral games, Hahm, Hilpert, and König (2024) presented respondents a short background information overview and instructions. For the dictator game, these were: *This game is played by pairs of individuals. Each pair is made up of a Player 1 and a Player 2. Each player will have some information about the other player, but you will not be told who the other players are during or after the experiment. The game is conducted as follows: A sum of 10 tokens will be provisionally allocated to Player 1 at the start of each round. Player 1 will then decide how much of the 10 tokens to offer to Player 2. Player 1 could give some, all, or none of the 10 tokens. Player 1 keeps all tokens not given to Player 2. Player 2 gets to keep all the tokens Player 1 offers. You will play this game three times with three different people.* In the trust game, the provided information

and instruction were: *This game is played by pairs of individuals. Each pair is made up of a Player 1 and a Player 2. Each player will have some information about the other player, but you will not be told who the other players are during or after the experiment. Each player will receive 10 tokens. Player 1 then has the opportunity to give a portion of his or her 10 tokens to Player 2. Player 1 could give some, all, or none of the 10 tokens. Whatever amount Player 1 decides to give to Player 2 will be tripled before it is passed on to Player 2. Player 2 then has the option of returning any portion of this tripled amount to Player 1. Then, the game is over. Player 1 receives whatever he or she keeps from the original 10 tokens, plus anything returned to him or her by Player 2. Player 2 receives their original 10 tokens, plus whatever he or she keeps after returning any portion of the tripled amount to Player 1. You will play this game three times, with three different people. The more tokens you obtain, the more successful you will be.*

In both games respondents were shown a tabular overview of Player 2 after the instructions. Figure 9 shows an example of such a profile along with the interface respondents were provided to assign the 10 tokens. Each round, a new profile was displayed to respondents.

	Player 2
Nationality	United Kingdom
Age	18
Party Affiliation	Labour Party (Labour)
Gender	Female
Religion	Muslim
Social Class	Middle Class

So put the number of tokens you wish to keep in the box labeled "Player 1." Put the tokens you wish to go to Player 2 in the box labeled "Player 2."

Player 1 (You)	0	Token(s)
Player 2	0	Token(s)
Total	0	Token(s)

Figure 9: Example of potential co-player profile.

8.3 Distribution of Y

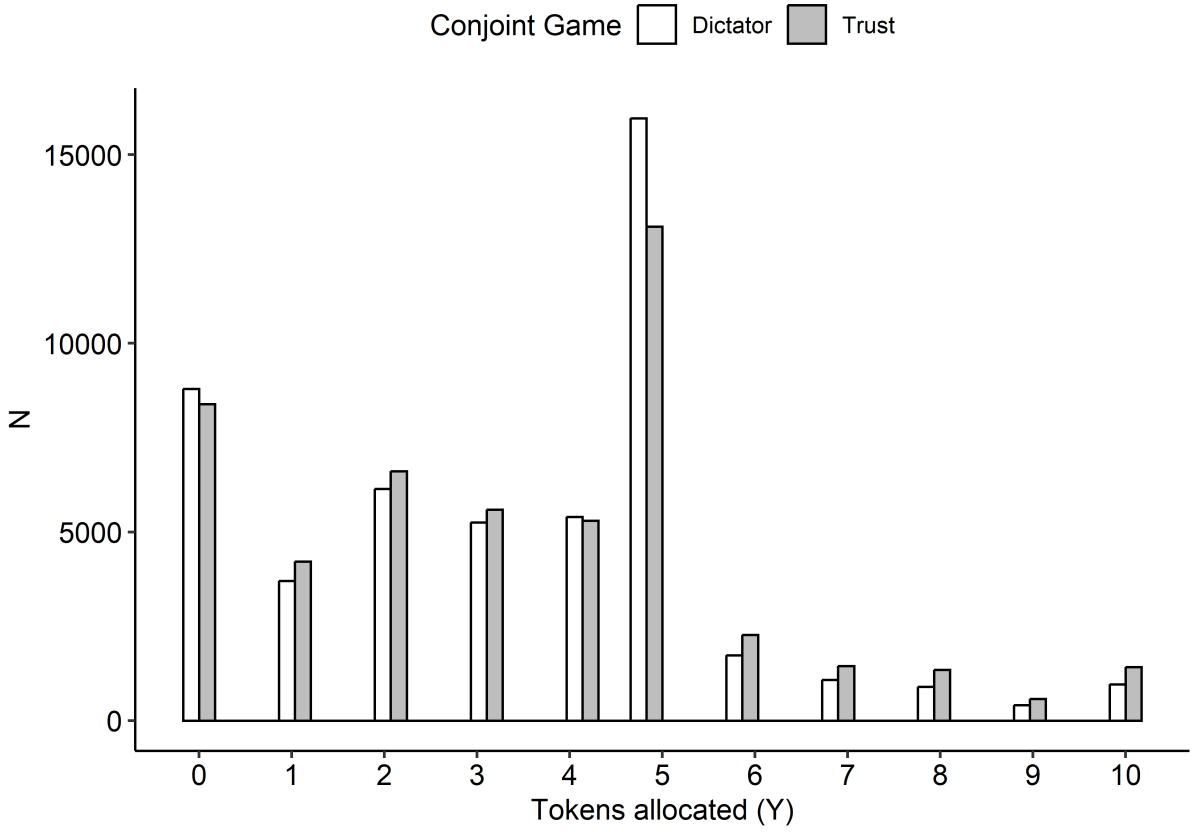

Figure 10: Distribution of token allocation (Y) by game. Dictator game: $Mean = 3.41$, $median = 4$, $SD = 2.35$. Trust game: $Mean = 3.48$, $median = 4$, $SD = 2.49$

8.4 Distribution of T and R

Table 3

der_partisan_type	None	Co	Out
0	5123	1948	18465
1	14898	30820	29366

8.5 Distribution of Covariates by T

Table 4

Variable	**0** N = 25,536[1]	**1** N = 75,084[1]
q_lrpos2_z	-0.08 (-0.45, 0.29)	-0.08 (-0.82, 0.66)
q_eupos2_z	0.08 (-0.68, 0.46)	0.08 (-0.68, 0.84)
q_econ_nativism_z	0.15 (-1.03, 0.74)	0.15 (-1.03, 0.74)
q_cult_nativism_z	0.07 (-1.03, 0.62)	0.07 (-1.03, 0.62)
q_satis_demo_country_z	0.30 (-0.81, 1.40)	0.30 (-0.81, 0.30)

Table 4

Variable	0 N = 25,536[1]	1 N = 75,084[1]
q_understand_nat_pol_z	0.12 (-0.65, 0.12)	0.12 (-0.65, 0.90)
q_understand_eu_pol_z	0.22 (-0.50, 0.22)	0.22 (-0.50, 0.94)
q_parties_harm_z	0.25 (-0.38, 0.88)	0.25 (-0.38, 0.88)
q_officials_talk_action_z	0.41 (-0.33, 1.15)	0.41 (-0.33, 1.15)
q_politics_good_evil_z	-0.14 (-0.76, 0.48)	-0.14 (-0.76, 0.48)
q_people_unaware_z	0.40 (-0.81, 1.00)	-0.20 (-0.81, 1.00)
q_leaders_educated_z	0.39 (-0.32, 1.10)	0.39 (-1.03, 1.10)
q_expert_decisions_z	0.16 (-0.49, 0.80)	0.16 (-0.49, 0.80)
q_listen_other_groups_z	0.18 (-0.74, 1.10)	0.18 (-0.74, 1.10)
q_democracy_compromise_z	-0.29 (-0.29, 0.57)	-0.29 (-0.29, 0.57)
q_interest_pol_country_z	0.06 (-0.60, 0.72)	0.06 (-0.60, 0.72)
q_interest_pol_eu_z	-0.28 (-0.96, 0.39)	0.39 (-0.28, 1.06)
q_eval_finance_household_z	0.01 (-0.99, 1.00)	0.01 (-0.99, 1.00)
q_eval_job_z	0.11 (-0.80, 1.03)	0.11 (-0.80, 0.11)
q_eval_econ_country_z	-0.14 (-1.03, 0.76)	-0.14 (-1.03, 0.76)
q_eval_econ_eur_z	0.07 (-0.94, 1.09)	0.07 (-0.94, 1.09)
q_risk_taking_z	0.11 (-0.56, 0.78)	0.11 (-0.56, 0.78)
q_future_discount_z	-0.17 (-0.84, 0.50)	-0.17 (-0.84, 0.50)
q_edu_z	-0.08 (-0.73, 0.57)	-0.08 (-0.73, 0.57)
q_age_z	-0.11 (-0.92, 0.64)	0.09 (-0.79, 0.91)
q_religion_en		
catholic	8,606 (34%)	27,379 (36%)
no religion	9,462 (37%)	25,678 (34%)
protstnt	1,881 (7.4%)	7,570 (10%)
other religion	5,470 (21%)	13,791 (18%)
muslim	114 (0.4%)	664 (0.9%)
q_perc_class		
Working class	5,307 (22%)	15,198 (21%)
Lower middle class	4,866 (20%)	13,924 (19%)
Middle class	12,060 (49%)	35,019 (48%)
Upper middle class	2,055 (8.4%)	7,622 (10%)
Upper class	191 (0.8%)	1,310 (1.8%)
q_rural_urban		
Rural area or village	5,819 (23%)	17,093 (23%)
Small or middle sized town	9,069 (36%)	28,418 (38%)
Large town	10,573 (42%)	29,298 (39%)
q_gender		
Male	10,939 (43%)	39,456 (53%)
Female	14,565 (57%)	35,514 (47%)
Other	32 (0.1%)	114 (0.2%)

[1] Median (Q1, Q3); n (%)

8.6 Robustness

8.7 Model summaries

```
> summary(bfit_cov_d)
Warning message:
There were 2 divergent transitions after warmup. Increasing adapt_delta above 0.8 may help. See http://mc-stan.org/misc/warnings.html#divergent-transitions-after-warm
 Family: gaussian
  Links: mu = identity
Formula: cj_token ~ 1 + der_partisan_type + der_partisan_relationship + der_partisan_type * der_partisan_relationship + cj_age_en + cj_class_en + cj_sex_en + cj_reli
   Data: filter(df_modelvars, meta_game == "dict") (Number of observations: 38240)
  Draws: 4 chains, each with iter = 5000; warmup = 3500; thin = 1;
         total post-warmup draws = 6000

Multilevel Hyperparameters:
~meta_country (Number of levels: 25)
                                                    Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)                                           0.16      0.04     0.08     0.26 1.00     2222     2350
sd(der_partisan_type)                                   0.05      0.04     0.00     0.14 1.01     1134     2373
sd(der_partisan_relationshipCo)                         0.08      0.05     0.01     0.19 1.00     1286     2177
sd(der_partisan_relationshipOut)                        0.11      0.05     0.02     0.21 1.00     1104     1236
cor(Intercept,der_partisan_type)                        0.08      0.43    -0.74     0.83 1.00     4580     4130
cor(Intercept,der_partisan_relationshipCo)             -0.04      0.38    -0.73     0.72 1.00     3858     4390
cor(der_partisan_type,der_partisan_relationshipCo)     -0.08      0.44    -0.82     0.76 1.00     2176     4056
cor(Intercept,der_partisan_relationshipOut)            -0.22      0.34    -0.77     0.53 1.00     2722     3023
cor(der_partisan_type,der_partisan_relationshipOut)     0.01      0.43    -0.79     0.79 1.00     1330     2686
cor(der_partisan_relationshipCo,der_partisan_relationshipOut)    -0.10      0.41    -0.83     0.68 1.00     1705     2922

~meta_country:der_partisan_anchor (Number of levels: 372)
                                                    Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)                                           0.05      0.03     0.00     0.13 1.00     1618     2605
sd(der_partisan_type)                                   0.06      0.04     0.00     0.16 1.01      580     1834
sd(der_partisan_relationshipCo)                         0.16      0.05     0.06     0.25 1.01      692     1126
sd(der_partisan_relationshipOut)                        0.15      0.04     0.08     0.22 1.00      648     1324
cor(Intercept,der_partisan_type)                       -0.15      0.46    -0.88     0.76 1.00     1010     2258
cor(Intercept,der_partisan_relationshipCo)              0.10      0.42    -0.73     0.82 1.01      392      738
cor(der_partisan_type,der_partisan_relationshipCo)     -0.08      0.41    -0.79     0.75 1.01      487     1139
cor(Intercept,der_partisan_relationshipOut)            -0.03      0.44    -0.80     0.79 1.02      247      748
cor(der_partisan_type,der_partisan_relationshipOut)     0.16      0.39    -0.63     0.84 1.01      400      931
cor(der_partisan_relationshipCo,der_partisan_relationshipOut)     0.10      0.31    -0.57     0.63 1.01      749     1733

~meta_pid (Number of levels: 18142)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     1.41      0.01     1.38     1.44 1.00     1853     3054

Regression Coefficients:
                                          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept                                     3.82      0.10     3.63     4.01 1.00     3117     4327
der_partisan_type                             0.01      0.06    -0.12     0.12 1.00     2847     3740
der_partisan_relationshipCo                   0.45      0.10     0.26     0.64 1.00     3198     3716
der_partisan_relationshipOut                 -0.48      0.06    -0.60    -0.35 1.00     3326     3846
cj_age_en30                                   0.09      0.03     0.03     0.16 1.00     4269     4299
cj_age_en42                                   0.09      0.03     0.02     0.15 1.00     3920     4967
cj_age_en53                                   0.08      0.03     0.02     0.15 1.00     4157     4678
cj_age_en65                                   0.05      0.03    -0.01     0.11 1.00     4010     4219
cj_class_en2_low_class                       -0.03      0.03    -0.08     0.02 1.00     6013     4958
cj_class_en3_high_class                      -0.15      0.03    -0.20    -0.10 1.00     6008     4686
cj_sex_en2_male                              -0.07      0.02    -0.11    -0.03 1.00     7335     5089
cj_reli_en2_catholic                         -0.06      0.03    -0.11     0.00 1.00     5269     4773
cj_reli_en3_protestant                       -0.15      0.03    -0.21    -0.09 1.00     5033     4654
cj_reli_en4_muslim                           -0.60      0.03    -0.66    -0.55 1.00     5148     4447
cj_eupos_shown                               -0.07      0.03    -0.12    -0.02 1.00     6785     4736
meta_round2                                   0.02      0.02    -0.03     0.06 1.00     7217     5160
meta_round3                                   0.10      0.02     0.05     0.14 1.00     7282     4561
meta_wave                                     0.02      0.03    -0.04     0.08 1.00     3530     3915
q_lrpos2_z                                   -0.07      0.02    -0.10    -0.03 1.00     3431     4583
q_eupos2_z                                    0.09      0.02     0.06     0.13 1.00     3445     4103
q_econ_nativism_z                             0.01      0.02    -0.04     0.06 1.00     3104     4239
q_cult_nativism_z                            -0.04      0.02    -0.09     0.00 1.00     3079     3851
q_religion_ennoreligion                      -0.12      0.04    -0.20    -0.05 1.00     3041     4119
q_religion_enprotstnt                        -0.08      0.06    -0.20     0.04 1.00     3210     4060
q_religion_enotherreligion                   -0.02      0.05    -0.12     0.07 1.00     3148     3767
q_religion_enmuslim                           0.15      0.17    -0.19     0.48 1.00     3294     4268
q_satis_demo_country_z                       -0.05      0.02    -0.09    -0.02 1.00     2944     3946
q_understand_nat_pol_z                       -0.08      0.02    -0.12    -0.04 1.00     3274     4243
q_understand_eu_pol_z                        -0.02      0.02    -0.06     0.01 1.00     3088     3747
q_parties_harm_z                             -0.02      0.02    -0.06     0.02 1.00     3234     4185
q_officials_talk_action_z                    -0.05      0.02    -0.09    -0.02 1.00     3313     4115
q_politics_good_evil_z                        0.03      0.02     0.00     0.06 1.00     3501     4341
q_people_unaware_z                           -0.01      0.02    -0.04     0.03 1.00     3161     3800
q_leaders_educated_z                         -0.09      0.02    -0.13    -0.06 1.00     3010     3740
q_expert_decisions_z                          0.06      0.02     0.03     0.10 1.00     3458     4393
q_listen_other_groups_z                      -0.02      0.02    -0.06     0.02 1.00     3315     4305
q_democracy_compromise_z                      0.02      0.02    -0.02     0.06 1.00     3380     4198
q_interest_pol_country_z                     -0.08      0.02    -0.13    -0.04 1.00     2929     4020
q_interest_pol_eu_z                           0.11      0.02     0.07     0.16 1.00     2971     4051
q_edu_z                                      -0.01      0.02    -0.04     0.02 1.00     3518     4168
q_perc_classLowermiddleclass                 -0.09      0.05    -0.18     0.00 1.00     2966     4099
q_perc_classMiddleclass                      -0.09      0.04    -0.17    -0.02 1.00     2722     3527
q_perc_classUppermiddleclass                 -0.19      0.06    -0.30    -0.07 1.00     2769     3539
q_perc_classUpperclass                       -0.35      0.12    -0.58    -0.13 1.00     3342     4489
q_eval_finance_household_z                    0.01      0.02    -0.04     0.05 1.00     2834     3654
q_eval_job_z                                 -0.03      0.02    -0.06     0.01 1.00     3100     4097
q_eval_econ_country_z                        -0.04      0.02    -0.07    -0.00 1.00     3343     4439
q_eval_econ_eur_z                             0.01      0.02    -0.03     0.04 1.00     3370     4321
q_genderFemale                               -0.06      0.03    -0.12    -0.00 1.00     3336     4106
q_genderOther                                 0.19      0.40    -0.60     0.98 1.00     3889     4328
q_age_z                                      -0.16      0.02    -0.19    -0.13 1.00     3417     3580
q_rural_urbanSmallormiddlesizedtown           0.01      0.04    -0.06     0.09 1.00     3107     3734
```

```
q_rural_urbanLargetown                          -0.00      0.04    -0.08     0.07 1.00     3367     3925
q_risk_taking_z                                  0.04      0.02     0.01     0.07 1.00     3656     4337
q_future_discount_z                              0.04      0.02     0.01     0.07 1.00     3364     4050
der_partisan_type:der_partisan_relationshipCo    0.09      0.10    -0.10     0.29 1.00     3108     3867
der_partisan_type:der_partisan_relationshipOut   0.03      0.06    -0.10     0.16 1.00     3282     4009

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     1.74      0.01     1.73     1.76 1.00     2917     4009

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).



> summary(bfit_cov_t)
 Family: gaussian
  Links: mu = identity
Formula: cj_token ~ 1 + der_partisan_type + der_partisan_relationship + der_partisan_type * der_partisan_relationship + cj_age_en + cj_class_en + cj_sex_en + cj_reli
   Data: filter(df_modelvars, meta_game == "trust") (Number of observations: 38222)
  Draws: 4 chains, each with iter = 5000; warmup = 3500; thin = 1;
         total post-warmup draws = 6000

Multilevel Hyperparameters:
~meta_country (Number of levels: 25)
                                                    Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)                                           0.14      0.05     0.03     0.24 1.01      644      349
sd(der_partisan_type)                                   0.09      0.05     0.00     0.20 1.00      566     1630
sd(der_partisan_relationshipCo)                         0.09      0.05     0.01     0.20 1.01      963     1888
sd(der_partisan_relationshipOut)                        0.04      0.03     0.00     0.11 1.00     2053     2325
cor(Intercept,der_partisan_type)                       -0.08      0.41    -0.78     0.75 1.00     2006     3058
cor(Intercept,der_partisan_relationshipCo)             -0.21      0.39    -0.84     0.64 1.00     2186     3221
cor(der_partisan_type,der_partisan_relationshipCo)     -0.23      0.43    -0.88     0.68 1.00     1585     2755
cor(Intercept,der_partisan_relationshipOut)            -0.03      0.44    -0.81     0.79 1.00     4695     3964
cor(der_partisan_type,der_partisan_relationshipOut)     0.02      0.45    -0.82     0.83 1.00     4005     4544
cor(der_partisan_relationshipCo,der_partisan_relationshipOut)  0.03  0.44    -0.79     0.81 1.00     4428     5014

~meta_country:der_partisan_anchor (Number of levels: 371)
                                                    Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)                                           0.13      0.06     0.01     0.24 1.02      394      881
sd(der_partisan_type)                                   0.09      0.06     0.00     0.22 1.01      348      561
sd(der_partisan_relationshipCo)                         0.28      0.05     0.18     0.36 1.01      578     1309
sd(der_partisan_relationshipOut)                        0.11      0.05     0.01     0.21 1.01      453      872
cor(Intercept,der_partisan_type)                       -0.08      0.44    -0.82     0.78 1.00      983     2076
cor(Intercept,der_partisan_relationshipCo)             -0.08      0.35    -0.71     0.68 1.02      136      247
cor(der_partisan_type,der_partisan_relationshipCo)     -0.27      0.40    -0.87     0.63 1.01      166      417
cor(Intercept,der_partisan_relationshipOut)             0.01      0.39    -0.69     0.78 1.00      984     2081
cor(der_partisan_type,der_partisan_relationshipOut)    -0.10      0.43    -0.83     0.76 1.00      507     1602
cor(der_partisan_relationshipCo,der_partisan_relationshipOut)  -0.15  0.34  -0.80     0.50 1.00      910     2170

~meta_pid (Number of levels: 18165)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     1.59      0.01     1.56     1.62 1.00     1784     3258

Regression Coefficients:
                                          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept                                     3.97      0.10     3.78     4.17 1.00     2426     3857
der_partisan_type                             0.05      0.07    -0.09     0.18 1.00     1888     3041
der_partisan_relationshipCo                   0.74      0.11     0.53     0.95 1.00     2293     3494
der_partisan_relationshipOut                 -0.33      0.06    -0.45    -0.21 1.00     2268     3707
cj_age_en30                                   0.08      0.04     0.01     0.15 1.00     3139     4331
cj_age_en42                                   0.14      0.03     0.07     0.20 1.00     3137     4344
cj_age_en53                                   0.07      0.03     0.01     0.14 1.00     3227     4407
cj_age_en65                                   0.01      0.04    -0.06     0.08 1.00     3131     3720
cj_class_en2_low_class                       -0.03      0.03    -0.08     0.02 1.00     4084     4348
cj_class_en3_high_class                      -0.16      0.03    -0.22    -0.11 1.00     3463     3574
cj_sex_en2_male                              -0.10      0.02    -0.14    -0.06 1.00     4513     4366
cj_reli_en2_catholic                         -0.01      0.03    -0.07     0.05 1.00     3403     3960
cj_reli_en3_protestant                       -0.10      0.03    -0.16    -0.04 1.00     3283     4137
cj_reli_en4_muslim                           -0.65      0.03    -0.71    -0.59 1.00     3085     4099
cj_eupos_shown                               -0.12      0.03    -0.17    -0.07 1.00     4926     4149
meta_round2                                  -0.02      0.02    -0.06     0.03 1.00     5395     4580
meta_round3                                   0.02      0.02    -0.03     0.06 1.00     5366     4530
meta_wave                                     0.03      0.03    -0.04     0.09 1.00     2272     4021
q_lrpos2_z                                   -0.04      0.02    -0.08    -0.01 1.00     2051     3205
q_eupos2_z                                    0.07      0.02     0.03     0.11 1.00     2454     3639
q_econ_nativism_z                             0.03      0.03    -0.02     0.09 1.01     1922     2904
q_cult_nativism_z                            -0.07      0.03    -0.13    -0.02 1.00     1961     2878
q_religion_ennoreligion                      -0.07      0.04    -0.14     0.01 1.00     2025     2619
q_religion_enprotstnt                        -0.03      0.07    -0.15     0.10 1.00     2009     3234
q_religion_enotherreligion                   -0.06      0.06    -0.17     0.05 1.01     1655     2797
q_religion_enmuslim                           0.20      0.19    -0.16     0.56 1.00     2399     3519
q_satis_demo_country_z                       -0.09      0.02    -0.12    -0.05 1.00     2350     3623
q_understand_nat_pol_z                       -0.10      0.02    -0.13    -0.06 1.00     2420     3481
q_understand_eu_pol_z                        -0.04      0.02    -0.08    -0.00 1.00     2164     3416
q_parties_harm_z                             -0.01      0.02    -0.05     0.02 1.00     2401     3948
q_officials_talk_action_z                    -0.06      0.02    -0.10    -0.02 1.00     2405     3542
q_politics_good_evil_z                        0.03      0.02    -0.00     0.07 1.00     2356     3706
q_people_unaware_z                            0.01      0.02    -0.03     0.04 1.00     2349     3350
q_leaders_educated_z                         -0.07      0.02    -0.11    -0.03 1.00     2438     3441
q_expert_decisions_z                          0.08      0.02     0.04     0.11 1.00     2611     3444
q_listen_other_groups_z                      -0.02      0.02    -0.06     0.02 1.00     2544     3492
q_democracy_compromise_z                      0.02      0.02    -0.02     0.06 1.00     2408     3596
q_interest_pol_country_z                     -0.07      0.03    -0.12    -0.02 1.00     2363     3698
q_interest_pol_eu_z                           0.11      0.03     0.06     0.16 1.00     2409     3529
q_edu_z                                       0.03      0.02    -0.01     0.06 1.00     2275     3179
q_perc_classLowermiddleclass                 -0.10      0.05    -0.20    -0.01 1.00     2003     3571
q_perc_classMiddleclass                      -0.08      0.04    -0.16     0.01 1.00     1657     3409
q_perc_classUppermiddleclass                 -0.06      0.06    -0.18     0.07 1.00     1729     2715
q_perc_classUpperclass                       -0.22      0.13    -0.47     0.03 1.00     2240     3758
q_eval_finance_household_z                   -0.00      0.02    -0.04     0.04 1.00     2023     3187
q_eval_job_z                                 -0.00      0.02    -0.04     0.04 1.00     2278     3941
q_eval_econ_country_z                        -0.05      0.02    -0.09    -0.01 1.00     1972     3582
```

```
q_eval_econ_eur_z                                  0.01      0.02    -0.02     0.05 1.00     2241     3759
q_genderFemale                                    -0.24      0.03    -0.30    -0.17 1.00     2615     3460
q_genderOther                                     -0.04      0.41    -0.85     0.76 1.00     2456     2967
q_age_z                                           -0.16      0.02    -0.19    -0.13 1.00     2470     3631
q_rural_urbanSmallormiddlesizedtown               -0.02      0.04    -0.10     0.06 1.00     2213     3505
q_rural_urbanLargetown                            -0.07      0.04    -0.15     0.02 1.00     2111     3597
q_risk_taking_z                                    0.05      0.02     0.01     0.08 1.00     2255     3409
q_future_discount_z                                0.05      0.02     0.02     0.09 1.00     2441     3034
der_partisan_type:der_partisan_relationshipCo     -0.12      0.11    -0.32     0.09 1.00     2384     3561
der_partisan_type:der_partisan_relationshipOut    -0.12      0.07    -0.26     0.01 1.00     2285     3241

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     1.79      0.01     1.78     1.81 1.00     2420     3524

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Eidesstattliche Erklärung – Statutory Declaration

Hiermit versichere ich, dass diese Arbeit von mir persönlich verfasst ist und dass ich keinerlei fremde Hilfe in Anspruch genommen habe. Ebenso versichere ich, dass diese Arbeit oder Teile daraus weder von mir selbst noch von anderen als Leistungsnachweise andernorts eingereicht wurden. Wörtliche oder sinngemäße Übernahmen aus anderen Schriften und Veröffentlichungen in gedruckter oder elektronischer Form sind gekennzeichnet. Sämtliche Sekundärliteratur und sonstige Quellen sind nachgewiesen und in der Bibliographie aufgeführt. Das Gleiche gilt für graphische Darstellungen und Bilder sowie für alle Internet-Quellen. Ich bin ferner damit einverstanden, dass meine Arbeit zum Zwecke eines Plagiatsabgleichs in elektronischer Form anonymisiert versendet und gespeichert werden kann. Mir ist bekannt, dass von der Korrektur der Arbeit abgesehen und die Prüfungsleistung mit „nicht ausreichend“ bewertet werden kann, wenn die Erklärung nicht erteilt wird.

I hereby declare that the paper presented is my own work and that I have not called upon the help of a third party. In addition, I affirm that neither I nor anybody else has submitted this paper or parts of it to obtain credits elsewhere before. I have clearly marked and acknowledged all quotations or references that have been taken from the works of other. All secondary literature and other sources are marked and listed in the bibliography. The same applies to all charts, diagrams and illustrations as well as to all Internet sources. Moreover, I consent to my paper being electronically stores and sent anonymously in order to be checked for plagiarism. I am aware that the paper cannot be evaluated and may be graded “failed” (“nicht ausreichend”) if the declaration is not made.

Place, Date

Signature