Does identity matter in legitimacy judgments on algorithmic governance?*

David Karpa† Daria Gritsenko‡

August 8, 2025

Abstract

Despite the increasing deployment of algorithmic systems in public governance, these technologies face a persistent legitimacy deficit when compared to human decision-makers. Existing research has primarily focused on procedural features, transparency, and output performance to explain this phenomenon. Yet, the affective and identity-based mechanisms that underlie legitimacy judgments remain underexplored.

This paper addresses this gap by integrating theories of ecological rationality and affective polarization into the study of algorithmic legitimacy. Using a preregistered survey experiment conducted in Finland (N = 2040), we examine how partisan identities and group-based moral intuitions shape public evaluations of automated asylum authorization systems. We show that legitimacy judgments are strongly conditioned by affective polarization: design features that violate in-group moral expectations trigger sharply negative evaluations – what we term "moral red flags". The results indicate that legitimacy perceptions are not uniformly distributed but vary systematically with partisan identities, highlighting the value of disaggregated analysis in legitimacy research. By demonstrating that legitimacy is not a neutral assessment but a context-sensitive, identity-laden judgment, this study offers a novel perspective on how algorithmic governance is evaluated in polarized societies.

Keywords: algorithmic governance, legitimacy, affective polarization, moral heuristics, public administration, survey experiment

*This research was supported by the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (Grant agreement No. 101116772, project AGAPP – Algorithmic Governance: A Public Perspective).

†University of Helsinki

‡University of Helsinki

1 Introduction

Algorithmic governance has become central to how modern states administer complex domains, from policing to welfare, raising pressing questions about legitimacy (Katzenbach and Ulbricht 2019; Gritsenko and Wood 2022). While such automated systems can increase efficiency, their perceived legitimacy – whether citizens view them as rightful and fair – depends on more than performance. Survey evidence underscores this complexity: a multi-country study found only a slim majority of citizens back algorithmic decision-making, with support strongly shaped by demographic and partisan factors (Sidhu et al. 2024). Legitimacy theories emphasize that perceived justice and shared values underlie acceptance of authority (Grimmelikhuijsen and Meijer 2022). In the context of border control, where decisions touch on fundamental values of deservingness and solidarity, understanding what drives legitimacy perceptions is therefore crucial.

Social identities and emotional reactions are powerful mediators of legitimacy assessments (Tajfel 1979; Iyengar and Westwood 2015). Individuals tend to interpret policies through their group attachments and moral lenses. Moreover, from an ecological rationality standpoint, if algorithmic rules align with citizens' intuitive moral heuristics, outcomes may seem more just (Gigerenzer 2010; Gigerenzer and Todd 2012). Violations of deep-seated norms may provoke distrust and outrage. By linking these perspectives, this paper examines how partisan affect and identity cues influence legitimacy judgments of an automated asylum system in Finland. Our study therefore bridges emerging research on algorithmic public administration with theories of cognitive-affective evaluation, shedding light on the micro-foundations of algorithmic legitimacy.

2 Literature and Hypotheses

Automation can be defined as a "technology that actively selects data, transforms information, makes decisions, or controls processes" (Lee and See 2004). Therefore, automation can refer to a variety of processes, from rule-based execution of functions to self-regulating and

self-directing autonomous systems. AI systems refer to "a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments" (OECD 2019). AI systems are typically designed to operate with some degree of autonomy. Models for studying automation acceptance are therefore widely adopted in the acceptance of AI literature (Koenig 2024).

Mainstream work on public perception of AI in public policy and administration aims at clarifying the conditions under which citizens would accept more automation in decision-making and public service provision. One well-documented finding in this literature is a preference for so-called human-in-the-loop arrangements. By comparing automated decision-making (ADM) and human involvement, scholars found that ADM is perceived as less legitimate than human or hybrid decision-making for all dimensions of legitimacy (input, throughput, output) (Starke and Lünich 2020). ADM's legitimacy penalty is preserved even when human-led decisions are likely to result in significant errors (Martin and Waldman 2022). ADM is also perceived as less trustworthy, while lower in red tape, than human-led decisions (Ingrams et al. 2021). Human involvement is particularly welcomed in emotionally complex or pivotal cases (Martin and Waldman 2022; Yalcin et al. 2023). What makes individuals more favourable towards ADM is higher technical affinity and awareness of AI (Denk et al. 2022). ADM is also more accepted in cases where automated decisions align with personal preferences (Lünich and Kieslich 2024). This is especially true when ADM produces "positive outcomes", and in technical rather than ideologically-laden advisory roles (Haesevoets et al. 2024).

Scholars often engage in experimental studies of street-level decision-making to disentangle how specific properties of automated systems, in combination with demographic and attitudinal factors, contribute to automation appraisal or aversion. Grimmelikhuijsen (2023) tested the effect of transparency and explainability on perceived trustworthiness of automated street-level decision-making. They showed that explainability has a larger effect, and that the effect of transparency is context-dependent. Busch (2023) revealed that the more

individuals long for discretion in a situation at hand, the more negatively their attitudes towards ADM are. At the same time, people with higher trust in technology – which coincided with respondent's age – were more willing to accept ADM. Horvath et al. (2023) found that while most respondents prefer processes with more human involvement, decision accuracy and cost were even more important drivers of AI acceptance in permit applications. Individual factors—such as tendency to adopt technology, negatively correlated with age—acted as a moderator of AI acceptance. Scholars also explored the contextual embeddedness of ADM. They showed that trust in the deploying organization, as well as institutional and organizational structures into which algorithms are introduced, matter for public acceptance of street-level ADM (Schiff et al. 2023; Wenzelburger et al. 2024).

In the domain of public administration, algorithmic decision-making (ADM) is argued to pose novel challenges to government legitimacy, threatening input, throughput, and output legitimacy (Grimmelikhuijsen and Meijer 2022). As the studies reviewed above indicate, one challenge that has not been addressed in the extant literature is the social dimension of algorithmic legitimacy. Legitimacy has a social dimension to it, yet, it is often studied as either an individual belief or assuming that there is a normative alignment in society that provides a universal basis for legitimacy judgments. Since contemporary societies are characterized by polarization, there are shared values and lived experience on each end of the polarized spectrum. Hence, citizens' reactions to these challenges are likely filtered through partisan identities rather than neutral analysis. Political polarization is fundamentally about social group membership, not mere issue disagreement; any group affiliation tends to trigger in-group favoritism and out-group bias, casting politics as an "us-versus-them" dynamic (Iyengar and Krupenkin 2018; Druckman and Levendusky 2019). From the perspective of ecological rationality, people facing such complex issues rely on simple, ecologically adapted heuristics rather than exhaustive calculation (Gigerenzer and Todd 2012). Importantly, many of these heuristics are inherently social; for example, individuals often imitate peers or defer to in-group norms to gain acceptance (Gigerenzer and Gaissmaier 2011). In practice,

this means that citizens tend to regard an automated process as legitimate if they believe their social or partisan community endorses it. Sparse empirical evidence is consistent with this prediction: self-identified conservatives and progressives articulate systematically different rationales for trusting algorithmic governance (e.g., conservatives emphasize efficiency, progressives emphasize fairness) (Sidhu et al. 2024). Accounting for this suggests that strong preferences for human oversight in automated systems (Martin and Waldman 2022) may be more accurately interpreted as identity-affirming cues, rather than evidence of a generalized "algorithmic aversion" (Dietvorst et al. 2015). In sum, legitimacy judgments about AI and automation appear to track social-identity cues; people may effectively ask (subconsciously) "what would my group think?" when assessing algorithms. This suggests that polarized identities and identity-threat considerations will color public acceptance of automation as much as (if not more than) the objective merits of the technology or the need for regulation in a given context. After all, social identities in contemporary Western societies are predominantly influenced by partisanship (West and Iyengar 2022).

This study draws on the theory of ecological rationality (Gigerenzer and Todd 2012) which suggests that decision-making strategies are adjusted to the structure of the environment. More specifically, Gigerenzer and Gaissmaier (2011) shows in social contexts people rely on evolved intuitions and social cues that allow them to manage social situation and belong to a certain group. In other words, and in application to legitimacy, what is appropriate is what a person expects to be endorsed by the members of a group they consider themselves belonging to. We are the first study to explicitly address the study of both legitimacy and algorithmic governance from the perspective of ecological rationality. We find that there are *red flags* in citizens perceptions when it comes to algorithmic systems. For example, no accountability in the process for the decision of the system is a red flag for all our respondents. We also find – what we term – *moral red flags* that are strictly determined by partisan identities. Whereas for respondents on the left too high costs coupled with disadvantageous system settings for asylum-seekers poses a *moral red flag*, the same system

is actually preferred to an average system for those with right partisan identity. More than that, those on the right despise systems that are free of charge – even when asylum-seekers are (unlawfully) rejected anyway. We find that what it means to be a "legitimate" system, cannot be assessed in ignorance of social dynamics. We suggest that future studies should re-consider averaging preferences or simply "controlling" for policy preferences, because as we show here, the important findings lie precisely in the heterogeneity of partisan identities, that are so characterizing for contemporary societies.

Contributing to the body of existing literature on algorithmic governance, we are able to demonstrate that the "the line that divides personal taste and moral virtues" (Gigerenzer 2023, p.124) is indeed strictly socially determined when it comes to evaluating algorithmic travel authorization systems. Legitimacy evaluations of such systems are, in fact, predominantly moral decisions, particularly in policy domains that are highly moralized in the public discourse. We leverage the strong partisan divide in the highly moralized topic of migration policies (Brader et al. 2008; Dias and Lelkes 2022) to illustrate how group-based moral intuitions shape legitimacy perceptions. That moral behavior is a function of both mind and the environment – and that it reflects the structure of the social environment – is not new in cognitive science and psychology (Gigerenzer and Gaissmaier 2011; Ellemers et al. 2019), but adds an important new perspective in the study of regulatory governance. In this view, what is perceived as legitimate is not simply a result of procedural or outcome-based criteria, but rather of alignment with what individuals expect to be endorsed within the moral and normative boundaries of their in-group.

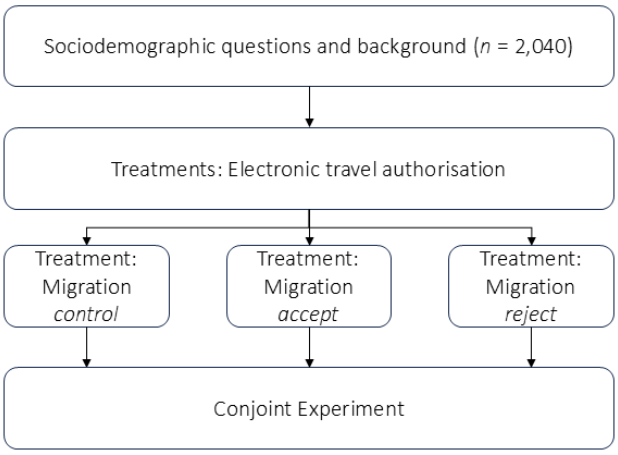
3 Data

In order to engage with these questions, we utilize a unique survey experiment conducted in October 2024. The second study is based on a preregistered survey that was carried out by Citizen Barometer based at the University of Helsinki.[1] The sample (N = 2040)

[1] https://aspredicted.org/48v2-9qk5.pdf
https://www.helsinki.fi/fi/projektit/kansalaisbarometri

consists of Finnish citizens over the age of 18. We asked participants questions about their sociodemographic background, knowledge on automated travel systems, and their attitudes towards automation in general. Representativeness weights were calculated based on respondents' language, region, age, gender, and education to align the sample with population benchmarks. We also asked participants about political opinions and voting behavior. We calculated the level of affective polarization by taking the distance between feeling thermometer scores towards political parties. The feeling thermometers asked participants how much they like a specific political party from *0 Strongly dislike* to *10 Strongly like*. For the analysis, we focused on the differences between the Perussuomalaiset (right-wing) and Vasemmistoliitto (left-wing) party, since they represent opposing poles of the political spectrum, while still being relevant for the public and political spheres. We concluded the survey with an experimental setup, in which participants were randomly assigned to one of three conditions, one of which served as a control group (henceforth *control*). In the other condition (*accept*), participants were told that the default setting of the system will be to accept all asylum seekers to the country. In the last condition (*reject*), participants were told that the default setting of the system will be to reject all asylum seekers' access to the country. In the *control* condition, the default of the system was simply not mentioned (see table A1). After being assigned to one of the three conditions (*accept, reject,* or *control*, participants were confronted with a classic conjoint experiment on automated systems in border control. In the conjoint experiment, participants saw two systems that differed on a range of randomly assigned attributes (Hainmueller et al. 2014). The conjoint experiment itself consisted of five attributes with three levels each (see table A2). Participants were asked to indicate which system they prefer and to rate both systems on a 5-point scale from very bad to very good. Participants had to choose three times (between two systems each) and thus rated six systems in total. The schematic flow of the experiment can be seen in Figure 1.

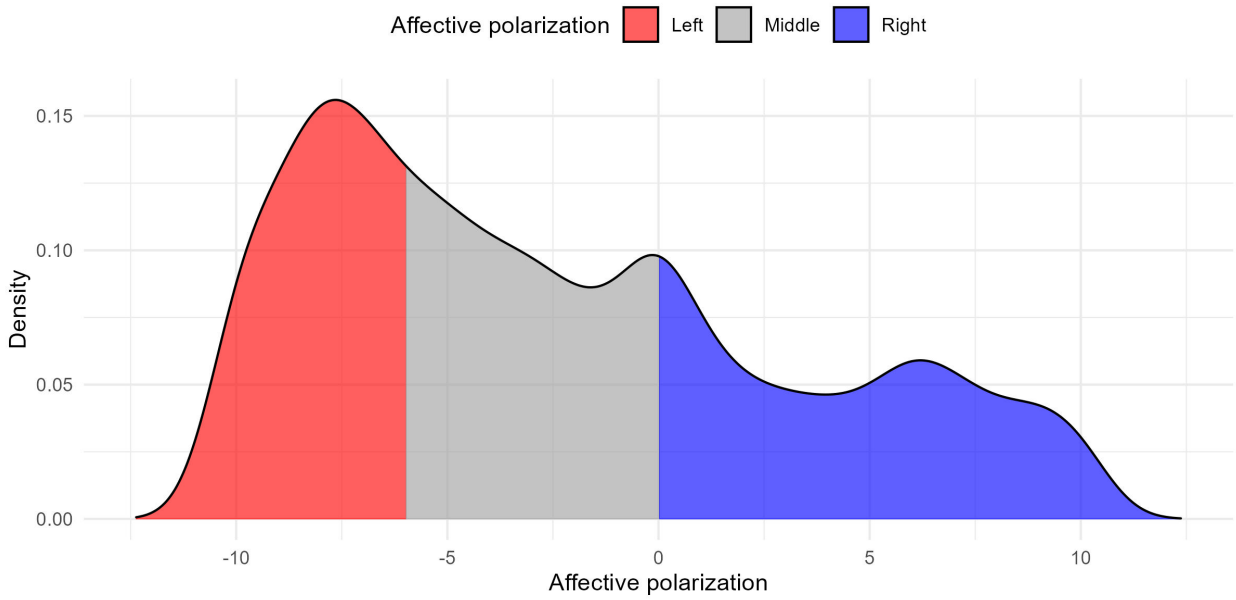
Figure 1: Experiment Flow



4 Results and Discussion

4.1 Descriptive Statistics

Figure 2: Distribution of Affective Polarization

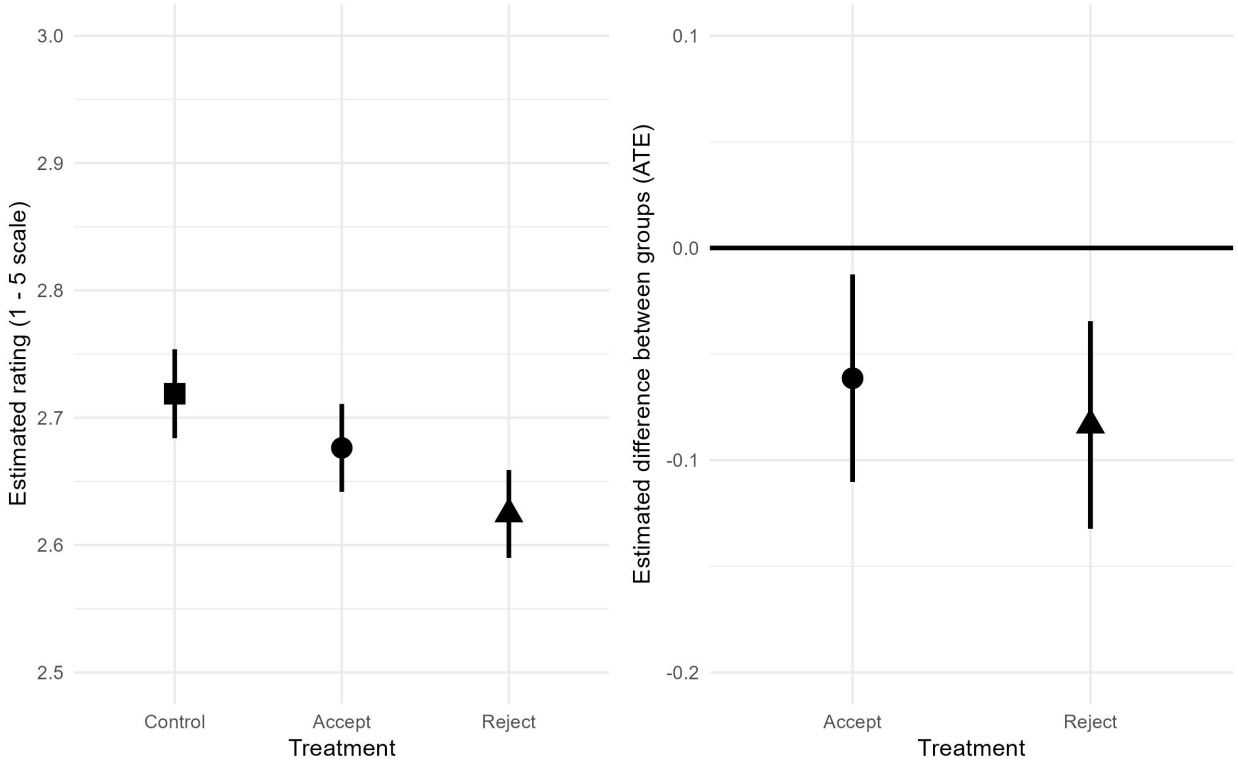


Note: This figure displays the distribution of affective polarization, operationalized as the absolute difference in feeling thermometer scores (0 to 10) between respondents' ratings of Perussuomalaiset (right-wing party) and Vasemmistoliitto (left-wing party). Higher values (i.e., values more distant from zero) indicate stronger partisan affect. Respondents were categorized into tertiles, corresponding to the lower ($\leq$33rd percentile; *Left*), middle (34th–66th percentile; *Middle*), and upper ($\geq$67th percentile; *Right*) thirds of the distribution.

Figure 2 presents the distribution of affective polarization among respondents. It illustrates the degree of partisan sentiment, which is central to the study's exploration of identity-driven legitimacy judgments.

4.2 Average Treatment Effects

Figure 3: Average Treatment Effects Across Experimental Conditions



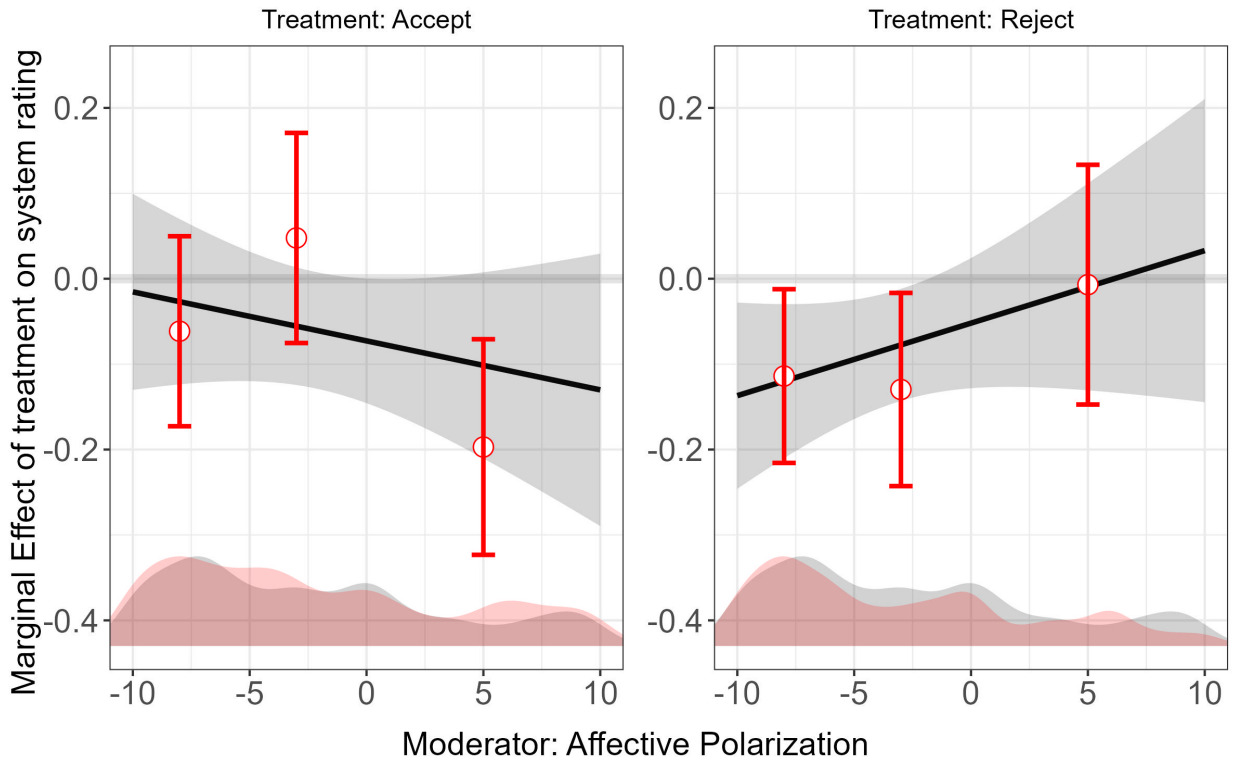
Note: This figure comprises two panels. The left panel displays predicted legitimacy ratings (on a 5-point Likert scale) across three treatment conditions: control, accept, and reject. These predictions are derived from an OLS regression model with robust standard errors, incorporating covariates for system attributes, affective polarization tertiles, and party preference variables. The right panel illustrates the estimated average treatment effects (ATEs) for the accept and reject conditions relative to the control group, with 95% confidence intervals.

Figure 3 displays the average effects of the experimental treatments on perceived legitimacy. The left panel of Figure 3 displays predicted legitimacy ratings across the three experimental conditions: *Control*, *Accept*, and *Reject*. The model incorporates covariates for system attributes, affective polarization, and party preferences. Participants in the Control group reported the highest average legitimacy rating (2.72), followed closely by the *Accept* con-

dition (2.68), and finally the *Reject* condition (2.62). The right panel of Figure 3 presents the estimated average treatment effects (ATEs) for the *Accept* and *Reject* conditions, each relative to the *Control* group. The *Accept* treatment is associated with a statistically significant decrease in perceived legitimacy of approximately -0.06 points ($p = 0.014$), while the *Reject* treatment shows a larger average decrease of about -0.08 points ($p < 0.001$). Both effects are statistically significant and fall below zero, indicating that introducing a default – whether inclusive or exclusive – reduces the perceived legitimacy of the system compared to the neutral condition in which no default is specified, on average.

4.3 Treatment Effects and Polarization

Figure 4: Conditional Average Treatment Effects by Polarization Level



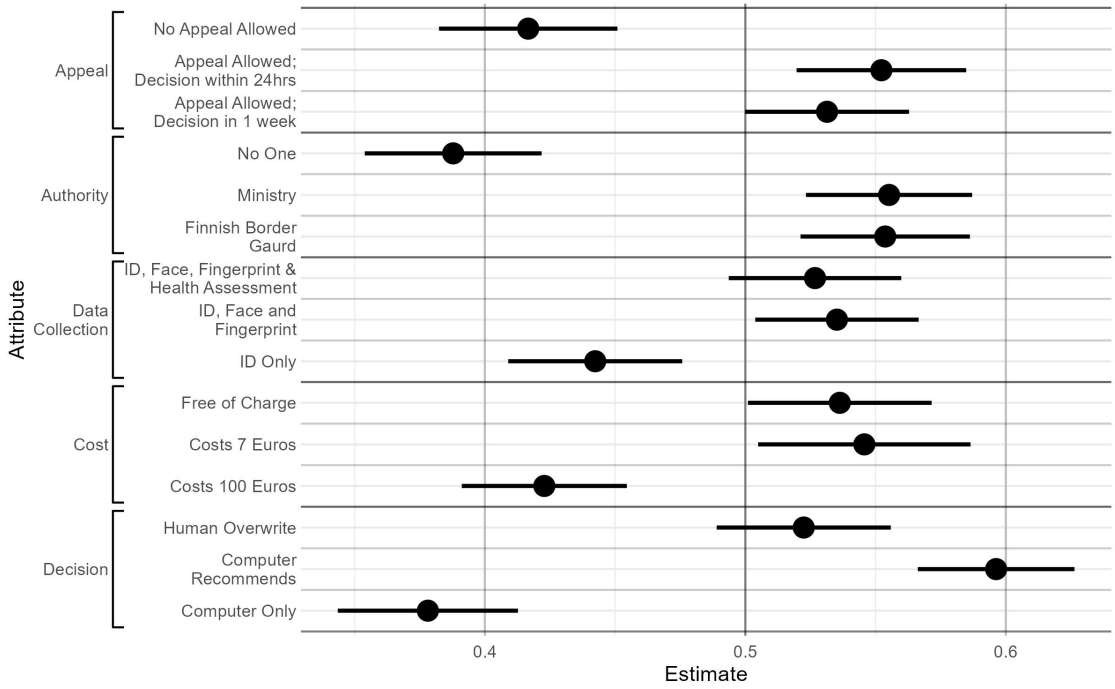
Note: This figure presents the marginal effects of treatment conditions (accept and reject) on legitimacy ratings, conditional on levels of affective polarization. Affective polarization is categorized into tertiles: left (≤33rd percentile), centre (34th–66th percentile), and right (≥67th percentile), see Figure 2. Estimates are obtained using the `interflex` package in R, employing a binning estimator with covariates for system attributes, party preference variables, and demographic controls. The shaded areas represent 95% confidence intervals.

Figure 4 explores how the treatment effects of algorithmic decision defaults vary across levels of polarization. The conditional average treatment effects (CATEs) are estimated using a binning estimator with covariates for system attributes, party preferences, and demographic controls. Essentially, we are trying to understand how an average individual rates an average system conditional on their partisan identity and the pre-conjoint treatment condition. We compare the *Accept* and *Reject* conditions with the *Control* group.

In the *Accept* - *Control* comparison, participants on the political right rate systems by about -0.2 points lower, whereas the the left and center groups rate the systems roughly the same (left panel). Similarly but with negative sign, the right panel shows the *Reject* - *Control* comparison. Here, participants on the left and in the center rate average systems worse, whereas the right is indifferent to system that reject asylum seekers per default (as compared to the control group). These patterns suggest that partisan intensity conditions how algorithmic defaults influence perceived legitimacy: Those who are on the right rate systems lower in which asylum-seekers are accepted and those who are on the left rate systems lower in which asylum-seekers are rejected. Notably, there is no statistically significant difference for both left and right when their preferred policy position is met.[2] The average treatment effects in Figure 3 are hence strongly moderated by partisan identity, such that they arise only when a policy-preference is violated for a respective group – and non-detectable otherwise.

[2] Dias and Lelkes (2022) show that migration is *the* diving policy when it comes to preferences and partisanship.

Figure 5: Predicted Attribute Preferences in the Control Group



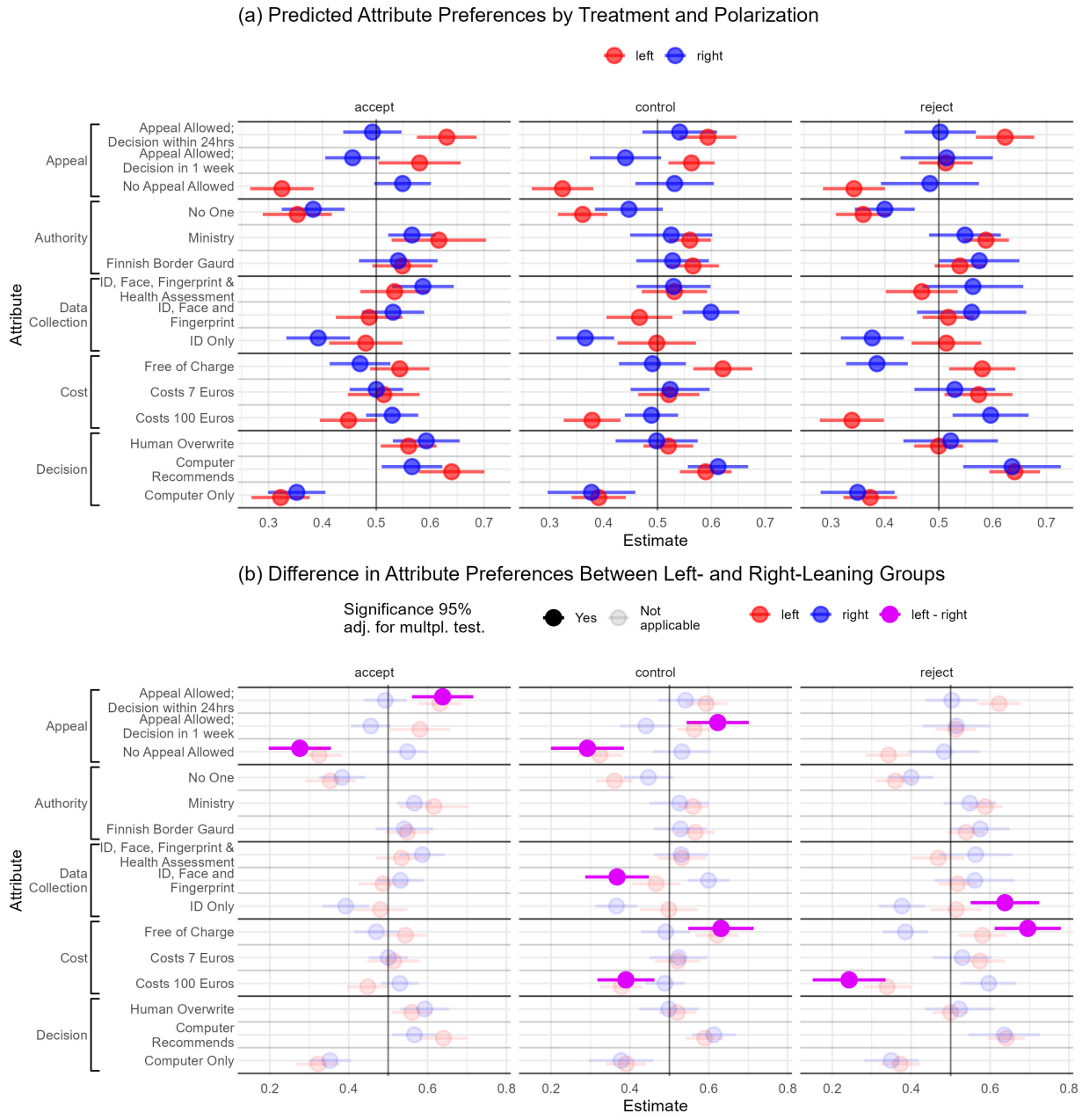
Note: This figure presents predicted evaluations of automated systems under the control condition using a multinomial logit model estimated via `cregg::mm()`. Estimates reflect the marginal effects in preferring an average system given a specific attribute level, with responses weighted for population representativeness. Deviations from 0.5 indicate positive or negative preferences, compared to indifference at 0.5. Confidence intervals denote 95% uncertainty.

Figure 5 shows predicted legitimacy ratings for different attribute levels of automated systems, estimated for the *Control* group only.[3] Respondents preferred systems when decisions were made with human oversight (*Computer Recommends*: 0.60, *Human Overwrite*: 0.52) compared to full automation (*Computer Only*: 0.38). Systems were also preferred when they were either low-cost or free to use, with *Costs 7 Euros* (0.55) and *Free of Charge* (0.54) preferred more than those costing *100 Euros* (0.42). In terms of data collection, preferences were highest for systems collecting *ID, Face and Fingerprint* (0.54) or also including a health assessment (0.53), relative to those relying on *ID Only* (0.44). Regarding the authority behind the decision, systems administered by the *Ministry* (0.56) or *Finnish Border Guard*

[3]We start by looking at only the *Control* group in order to understand preferences for algorithmic systems without experimental treatment effects. In terms of Figure 1, we are now looking at a third of respondents, that is, those in the left experimental arm labeled as *Control*.

(0.55) were preferred to those with *No One* being accountable (0.39). Finally, systems that permitted appeals – especially with faster processing (*within 24 hours*: 0.55) – were viewed more favorably than those offering no appeal process (0.42). These findings indicate that there is one *red flag* per attribute. Interestingly, costs play a big role, a finding that will show up in later analysis. This is surprising in light of the literature, where costs (especially non-substantial costs) are usually not found to matter in legitimacy evaluations. In order to assess whether *red flags* are treatment effects that show up unanimously for all participants, we turn to estimating marginal means for all treatments (*Control*, *Reject*, *Accept*) and for the polarized groups (*left*, *right*) at the poles of Figure 2.

Figure 6: Attribute Preferences by Treatment and Polarization



Note: Estimates are based on multinomial logit models stratified by treatment condition (control, accept, reject) and affective polarization group (left, right). Models are estimated using `cregg::mm()` with weighting applied. Estimates reflect the marginal effects in preferring an average system given a specific attribute level, with responses weighted for population representativeness. Deviations from 0.5 indicate positive or negative preferences, compared to indifference at 0.5. Panel (a) shows Marginal means estimated for all treatments (*Control*, *Reject*, *Accept*) and for the polarized groups (*left* in red, *right* in blue). Panel (b) shows differences between left- and right-leaning respondents which are computed using `cregg::mm_diffs()` after estimating group-specific multinomial logit models. The figure shows attribute-level contrasts across treatment conditions, with 95% confidence intervals corrected for multiple comparisons. Only confidence-intervals for contrasts (purple) are shown were the effect is statistically significant at the 95%-level after adjusting for multiple testing ($p * 15$, i.e., p-value multiplied by number of tests per condition). Transparent point estimates in blue and red in the background for right and left leaning groups (corresponding to (a)), respectively.

Panel (a) in Figure 6 presents predicted marginal means for different system attribute levels, disaggregated by treatment group and partisan orientation (*left* and *right*). Panel (b) shows differences between left- and right-leaning respondents as attribute-level contrasts across treatment conditions, with 95% confidence intervals corrected for multiple comparisons. Only confidence-intervals for contrasts (purple) are shown were the effect is statistically significant at the 95%-level after adjusting for multiple testing ($p*15$, i.e., p-value multiplied by number of tests per condition). Several patterns stand out. First, across all treatment arms, respondents on both the *left* and *right* consistently favor systems with some form of human involvement in decision-making – *Computer Recommends* and *Human Overwrite* – over *Computer Only* systems. We do not find meaningful differences in preferences for human involvement in the decision process, that is, human-in-the-loop preferences are *not* (measurably) moderated by partisan identity. Second, cost sensitivity differs by ideology: for left-leaning respondents, the legitimacy of *Free of Charge* systems remains high and stable, whereas among the right, cost-related evaluations shift more visibly, with sharp declines in legitimacy for both *Free* and *7 Euro* systems under the *Reject* treatment. The difference under the *Reject* and *Control* treatments shows that right-leaning respondents are significantly more favorable toward high-cost (*100 Euros*) systems, while left-leaning respondents continue to prefer low-cost and free systems. Third, right-leaning respondents show clearer preferences for expansive data collection practices (e.g., including health assessments), while the left assigns more consistent ratings across data regimes, with only modest variation across treatment. Panel (b) reveals that these differences in evaluations of data collection intensity are statistically significant: in the *Control* and *Reject* conditions, right-leaning respondents are significantly more supportive of privacy-invasive systems (e.g., *ID Only*) than their left-leaning counterparts. Fourth, in terms of institutional accountability, both groups consistently rate systems with formal oversight – especially the *Ministry* and *Border Guard* – as more legitimate than those lacking an identifiable decision-maker. There is no statistically detectable difference between the two groups. Finally, the availability of appeals

appears more consequential for the left: across treatments, appeal options (particularly with fast resolution) increase perceived legitimacy among left-leaning respondents more sharply than among the right. This difference is statistically significant. Across several treatments, we observe only muted partisan differences in preferences for decision-making authority and human oversight, suggesting a degree of cross-party consensus on institutional control and transparency. Together, these findings demonstrate that while some attribute preferences are ideologically stable, others – especially around cost, privacy, and recourse – become sharply polarized depending on the policy context.

The findings in Figure 6 can be interpreted through the lens of *moral red flags* – design features that violate core moral intuitions and therefore sharply lower the perceived legitimacy of algorithmic systems. Because migration policy is a moralized domain, individuals are unlikely to evaluate system designs solely based on procedural or instrumental considerations. Instead, they draw on normative commitments shaped by their *social environment*, particularly their position within partisan groups. Take the cost attribute for example – respondents on the left strongly disfavor systems that cost 100€ and reject asylum-seekers nevertheless. We argue that is a violation of a specific fairness principle, a *moral red flag*. Crucially, what is morally acceptable seems to be strictly determined by partisan identities.

Political polarization thus functions as a proxy for respondents' broader moral context: left- and right-leaning individuals inhabit distinct normative communities, each with its own expectations for what constitutes a fair, just, or legitimate asylum process. When a system exhibits features that conflict with these group-based norms – for instance, high costs for accessing services (a red flag for the left), or lack of sufficient identification from foreigners (a red flag for the right) – evaluations become strongly negative.

This framework helps explain why certain design features elicit asymmetric responses across partisan subgroups, and why these asymmetries become more pronounced in specific treatment conditions. For example, the absence of appeal mechanisms under the *Reject* condition becomes a moral red flag for left-leaning respondents, who interpret it as a violation

of fairness and procedural justice. In contrast, right-leaning respondents show greater cost sensitivity under the same condition, suggesting that financial thresholds may be morally justified as a form of filtering in the context of rejection.

In short, the observed heterogeneity in legitimacy evaluations reflects not just preference differences, but *moralized judgments* triggered by contextual cues. Treatment conditions activate different normative frames, and design features are interpreted accordingly – as morally acceptable, neutral, or suspect – depending on partisan alignment and moral orientation. We conclude that legitimacy judgments are inherently socially determined and address this gap in the literature.

5 Bibliography

Brader, T., Valentino, N. A., and Suhay, E. (2008). What triggers public opposition to immigration? anxiety, group cues, and immigration threat. *American Journal of Political Science*, 52(4):959–978.

Busch, P. A. (2023). Faced with digital bureaucrats: A scenario-based survey analysis of how clients perceive automation in street-level decision-making. *Government Information Quarterly*, 40(4):101872.

Denk, T., Hedström, K., and Karlsson, F. (2022). Citizens' attitudes towards automated decision-making. *Information Polity*, 27(3):391–408.

Dias, N. and Lelkes, Y. (2022). The nature of affective polarization: Disentangling policy disagreement from partisan identity. *American Journal of Political Science*, 66(3):775–790.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General*, 144(1):114.

Druckman, J. N. and Levendusky, M. S. (2019). What do we measure when we measure affective polarization? *Public Opinion Quarterly*, 83(1):114–122.

Ellemers, N., van der Toorn, J., Paunov, Y., and van Leeuwen, T. (2019). The psychology of morality: A review and analysis of empirical studies published from 1940 through 2017. *Personality and Social Psychology Review*, 23(4):332–366.

Gigerenzer, G. (2010). Moral satisficing: Rethinking moral behavior as bounded rationality. *Topics in cognitive science*, 2(3):528–554.

Gigerenzer, G. (2023). *The intelligence of intuition*. Cambridge University Press.

Gigerenzer, G. and Gaissmaier, W. (2011). Heuristic decision making. *Annual review of psychology*, 62(2011):451–482.

Gigerenzer, G. and Todd, P. (2012). *Ecological Rationality: Intelligence in the World*. Oxford University Press.

Grimmelikhuijsen, S. (2023). Explaining why the computer says no: Algorithmic trans-

parency affects the perceived trustworthiness of automated decision-making. *Public Administration Review*, 83(2):241–262.

Grimmelikhuijsen, S. and Meijer, A. (2022). Legitimacy of algorithmic decision-making: Six threats and the need for a calibrated institutional response. *Perspectives on Public Management and Governance*, 5(3):232–242.

Gritsenko, D. and Wood, M. (2022). Algorithmic governance: A modes of governance approach. *Regulation & Governance*, 16(1):45–62.

Haesevoets, T., Verschuere, B., Van Severen, R., and Roets, A. (2024). How do citizens perceive the use of artificial intelligence in public sector decisions? *Government Information Quarterly*, 41(1):101906.

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

Horvath, L., James, O., Banducci, S., and Beduschi, A. (2023). Citizens' acceptance of artificial intelligence in public services: Evidence from a conjoint experiment about processing permit applications. *Government Information Quarterly*, 40(4):101876.

Ingrams, A., Kaufmann, W., and Jacobs, D. (2021). In ai we trust? citizen perceptions of ai in government decision making. *Policy & Internet*, 14(2):390–409.

Iyengar, S. and Krupenkin, M. (2018). Partisanship as social identity; implications for the study of party polarization. *The Forum*, 16(1):23–45.

Iyengar, S. and Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*, 59(3):690–707.

Katzenbach, C. and Ulbricht, L. (2019). Algorithmic governance. *Internet Policy Review*, 8(4):1–18.

Koenig, P. D. (2024). Attitudes toward artificial intelligence: Combining three theoretical perspectives on technology acceptance. *AI & SOCIETY*.

Lee, J. D. and See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80.

Lünich, M. and Kieslich, K. (2024). Exploring the roles of trust and social group preference on the legitimacy of algorithmic decision-making vs. human decision-making for allocating covid-19 vaccinations. *AI & SOCIETY*, 39(1):309–327.

Martin, K. and Waldman, A. (2022). Governing algorithmic decisions: The role of decision importance and governance on perceived legitimacy of algorithmic decisions. *Big Data & Society*, 9(1):20539517221100449.

OECD (2019). Oecd recommendation on artificial intelligence (ai). Available at https://www.oecd.org/going-digital/ai/principles/.

Schiff, K. J., Schiff, D. S., Adams, I. T., McCrain, J., and Mourtgos, S. M. (2023). Institutional factors driving citizen perceptions of ai in government: Evidence from a survey experiment on policing. *Public Administration Review*.

Sidhu, D., Magistro, B., Allen Stevens, B., and Loewen, P. J. (2024). Why do citizens support algorithmic government? *Journal of Public Policy*, 44(3):659–677.

Starke, C. and Lünich, M. (2020). Artificial intelligence for political decision-making in the european union: Effects on citizens' perceptions of input, throughput, and output legitimacy. *Data & Policy*, 2:e16.

Tajfel, H. (1979). Individuals and groups in social psychology. *British Journal of Social and*

Clinical Psychology, 18(2):183–190.

Wenzelburger, G., König, P. D., Felfeli, J., and Achtziger, A. (2024). Algorithms in the public sector. why context matters. *Public Administration*, 102(1):40–60.

West, E. A. and Iyengar, S. (2022). Partisanship as a social identity: Implications for polarization. *Political Behavior*, 44(2):807–838.

Yalcin, G., Themeli, E., Stamhuis, E., Philipsen, S., and Puntoni, S. (2023). Perceptions of justice by algorithms. *Artificial Intelligence and Law*, 31(2):269–292.

Table A1: Automated travel systems treatment description

In the future, travelers coming to Finland from countries that do not need a visa to enter the EU will still need to get an authorisation 72h before their trip. This new system is called ETIAS, which stands for Electronic Travel Identification and Authorisation System. Individuals will apply for ETIAS online. Applications will be quickly processed by computer systems using personal data about travelers. On the next screen, you can see two alternatives of how ETIAS could be set up.		
The default setting of ETIAS will be to accept all asylum seekers to the country.	The default setting of ETIAS will be to prevent asylum seekers from entering the country.	[CONTROL]

A Appendix

A.1 Tables

Table A2: Conjoint Experiment Design

Attribute	**Levels**
1) Who makes entry decisions?	1a) Computer makes decisions without human involvement. 1b) Computer recommends, but a human makes the final decision. 1c) Computer makes decisions, but border guards can change the decision.
2) How much do applicants pay?	2a) Free to use. 2b) Costs EUR 7 per application. 2c) Costs EUR 100 per application.
3) Which kind of personal data will be collected?	3a) Collects ID, fingerprint, and face data. 3b) Collects ID, fingerprint, face and health self-assessment data. 3c) Collects ID.
4) Who is responsible when the system makes a mistake?	4a) The Finnish border guard (Rajavartiolaitos). 4b) Ministry of the Interior that developed the system. 4c) No one.
5) Is there a possibility to appeal?	5a) Appeals allowed, with a decision in 24 hours. 5b) Appeals allowed, with a decision in one week. 5c) The decision is final and cannot be appealed.

A.2 Figures