sociological science

Beyond Text: Using AI-Generated Visual Conjoints to Study Gender and Housework Attribution

Léa Pessin, Kevin Munger

European University Institute

Abstract: Despite substantial gender convergence in education and employment, women continue to perform a disproportionate share of housework. We employ a novel visual conjoint experiment to isolate the normative mechanisms underlying this persistent inequality. Using AI-generated photorealistic images, we systematically vary the tidiness of domestic spaces, room type, source of mess, socioeconomic status, and the gender and race/ethnicity of occupants, alongside text describing couples' employment arrangements. A quota sample of 2,994 U.S. respondents each evaluated five vignettes, yielding 14,970 observations. We find that gender effects operate primarily through responsibility attribution rather than through differential perception of messiness or anticipated social judgment. Women are assigned significantly more cleaning responsibility than men, with the gender penalty concentrated among dual-earner couples. Child-caused mess is perceived as messier than adult-caused mess yet carries reduced social consequences, suggesting that it operates as a legitimating excuse. Our findings suggest that gender equality in paid work is necessary for achieving gender equality in housework, but that it is not sufficient, and that this gap will persist absent changes in normative expectations around responsibility for housework.

Keywords: gender; housework; visual conjoint; experimental methods

Reproducibility Package: Replication materials including all data and code and documentation necessary to reproduce all empirical results reported in the article can be found at https://github.com/kmunger/Housekeeping_SocSci_Replication.

Citation: Pessin, Léa, Kevin Munger, 2026. "Beyond Text: Using AI-Generated Visual Conjoints to Study Gender and Housework Attribution" Sociological Science 13: 661-684.

Received: January 20, 2026

Accepted: April 14, 2026

Published: June 16, 2026

Editor(s): Ari Adut, Elizabeth Bruch

DOI: 10.15195/v13.a26

THE second half of the twentieth century has been characterized by dramatic changes in gender relations (Goldin 2014). Despite substantial gender convergence in college education attainment and labor force participation, progress toward gender equality in the domestic sphere and particularly in housework has been more modest, and slowing starting in the 1990s (Altintas and Sullivan 2016; England, Levine, and Mishel 2020; Pessin 2024). For example, in the United States, although the gender gap in housework decreased substantially over the last 70 years, today women in married couples still spend more than twice as much time as men in feminine core housework (defined as cleaning and laundry) (Milkie et al. 2025). This stubborn persistence of gender inequality in housework despite convergence in other domains represents a key puzzle that has motivated a large body of gender and family research (Ferree 2010; Ross-Plourde and Lachance-Grzela 2025; Sullivan 2021).

Scholars have proposed multiple explanations for this persistence (see Dominguez-Folgueras 2022, for a review). One strand of research emphasizes economic resources, arguing that housework allocation reflects bargaining power derived from relative earnings within couples (Bittman et al. 2003; Brines 1994).

A second approach focuses on time availability, suggesting that the partner with more time at home takes on more domestic work (Hook 2017). A third perspective highlights gender norms, examining how attitudes, identities, and gender performance shape household labor division (Risman 2004; Ridgeway and Correll 2004; West and Zimmerman 1987). However, as Dominguez-Folgueras (2022) notes, these mechanisms are not analytically independent: women's lower earnings partly result from anticipated domestic responsibilities, these responsibilities structure their time availability, time constraints reinforce gender norms about housework, and these norms in turn shape labor market outcomes. This interdependence poses significant challenges for observational research attempting to identify these mechanisms.

In this study, we employ experimental methods to isolate the gender normative dimension from these intertwined mechanisms. Following Thébaud, Kornrich, and Ruppanner (2021), we distinguish between two mechanisms through which gender norms may operate. At the individual level, a socialization account suggests that men and women internalize different standards for domestic order: women may literally "see" mess differently, or care more about tidiness due to gender socialization (Blau et al. 2020; Cunningham 2005; Pessin and Arpino 2018). In contrast, a social-interactional account posits that women are held to stricter standards by others, facing social sanctions and judgment that enforce gender role expectations (Berk 1985; Ridgeway and Correll 2004; West and Zimmerman 1987). Experimental designs that systematically vary both the state of the domestic space and the gender of the person being evaluated can distinguish between these mechanisms. Thébaud et al. (2021) provide pioneering evidence for these theoretical mechanisms through an experiment using photographic images of rooms that varied in tidiness and occupant gender, demonstrating that women face harsher judgment for identical levels of domestic disorder. The Thébaud et al. study is part of a nascent but growing body of research employing experimental methods to understand how gender shapes household dynamics, particularly regarding housework and the gender division of labor (Auspurg, Iacovou, and Nicoletti 2017b; Auspurg and Düval 2024; Carriero and Todesco 2017; Doan and Quadlin 2019; Doan, Quadlin, and Khanna 2024; Pedulla and Thébaud 2015; Pepin and Scarborough 2025; Schulz 2021; Urbina and Tisch 2025; Van Breeschoten, Roeters, and van der Lippe 2018).

We maintain the core theoretical logic of Thébaud et al. (2021), but extend their approach in several key dimensions, both methodologically and theoretically. In terms of methods, our study includes over four times the number of respondents and employs a multifactorial design where each respondent is shown five distinct vignettes, yielding a sample size roughly 22 times larger than the original.[1] To implement this design, we use generative AI to create images that systematically vary multiple dimensions while holding irrelevant factors constant. Substantively, we test for the importance of several novel factors beyond the original variation in room messiness and occupant's gender (operationalized through first names in their study). We vary room type (living room vs. kitchen), socioeconomic status (SES) (operationalized through furniture quality), and the source of mess (children's items vs. adult belongings). Another important innovation is that we signal the occupant gender through an AI-generated image instead of a first name, and vary their race/ethnicity as well. Finally, in addition to information about the employment

status of the occupant, we include information about the paid work status of their (different-sex) spouse.

We fielded our experiment using Prolific's quota sampling to approximate the U.S. population on gender, race/ethnicity, and age. Our final analytic sample consists of 2,994 respondents who each evaluated five vignettes, yielding 14,970 total observations. Each vignette presented respondents with an AI-generated image of a room, an AI-generated image of the room's occupant, and text describing the employment arrangement of the occupant and their spouse. Respondents answered three questions for each vignette: (1) how messy they perceived the room to be; (2) how they expected a visitor to judge the occupant based on the room's condition; and (3) who they believed was primarily responsible for cleaning the room. This design allows us to test both socialization mechanisms (do respondents perceive mess differently?) and social-interactional mechanisms (do they judge occupants differently and assign responsibility differently?). The experiment was preregistered.[2]

Our results show that gender norms operate primarily through responsibility attribution rather than perception or social judgment. Contrary to prior experimental work, we find no evidence that women's rooms are perceived as messier or that women face harsher anticipated social consequences for identical levels of mess. However, women are assigned significantly more cleaning responsibility than men, even after accounting for employment status. We also find that child-caused mess reduces social sanctions despite being perceived as messier, and that employment status shapes responsibility attribution, though traditional gender norms reassert themselves precisely in dual full-time arrangements where time availability would predict equality. We find limited evidence for effects of occupant race/ethnicity or SES.

Theoretical Mechanisms and Hypotheses

In this study, we focus on the role of gender norms in shaping inequality in the division of housework among different-sex couples. Building on prior work, we distinguish between individual-level and social-interactional theoretical mechanisms and use experimental methods to identify their effects. In what follows, we briefly review these theoretical mechanisms and articulate our hypotheses.

First, a socialization account suggests that men and women internalize different standards for domestic order through childhood experiences and ongoing gender socialization (Blau et al. 2020; Cunningham 2005; Pessin and Arpino 2018). From this perspective, women may literally perceive mess differently than men: they see rooms as messier because they have developed stronger preferences for tidiness. If this mechanism operates, we would expect female respondents to perceive rooms as messier than male respondents, holding constant the objective level of messiness and all occupant characteristics (**H1**).

Second, a social-interactional account posits that women are held to stricter standards by others, facing social sanctions and accountability pressures that enforce gender role expectations (Berk 1985; Risman 2004; Ridgeway and Correll 2004; West and Zimmerman 1987). This perspective emphasizes that individuals anticipate how others will judge them and adjust their behavior accordingly, regardless of their preferences. If this mechanism operates, we would expect rooms occupied

by women to elicit more negative anticipated social consequences than identical rooms occupied by men (**H2**), and women to be assigned greater responsibility for cleaning than men in identical domestic situations (**H3**).

Social-interactional accountability does not operate uniformly across all contexts. Rather, accountability practices are shaped by culturally resonant frames that vary depending on who is being evaluated, what space they occupy, and their situational circumstances (Ridgeway 2011; Risman 1998). The intersection of multiple social identities—gender, race, and class—may create distinct patterns of judgment, as individuals face compounded or contradictory expectations when it comes to housework and family life (Few-Demo and Allen 2020; McCall 2005; Pessin and Pojman 2025). Similarly, spatial contexts carry different levels of gendered meaning: kitchens are explicitly associated with feminine domestic labor, whereas living rooms function as more public, less clearly gendered spaces. The source of disorder also activates different cultural frames: children's mess may provide a legitimate excuse for domestic disorder, or it may intensify maternal accountability by making caregiving responsibilities more salient. Finally, employment arrangements signal time availability and normative expectations about who should maintain the home, yet these situational factors may be overridden by gender when accountability pressures are strong (Bittman et al. 2003; Greenstein 2000; Tichenor 2005).

We test whether these contextual dimensions independently affect accountability judgments and whether they moderate gender effects. For many of these interactions, competing theoretical mechanisms suggest different patterns. Race and ethnicity might intensify negative judgments of disorder through stereotype confirmation, or might reduce them if lower standards are applied to minority occupants; gender effects could be amplified for women of color who face intersecting pressures, or attenuated if racial stereotypes override gender expectations. Children's mess might reduce gender differences in responsibility attribution if it operates as a universally legitimating excuse, or it might amplify gender differences if it heightens specifically maternal accountability. Similarly, employment arrangements might attenuate gender effects when couples have egalitarian work patterns, or gender norms might persist even when women are primary breadwinners. By systematically varying these dimensions, we can identify the conditions under which gender-based accountability is amplified, attenuated, or complicated by other social factors. We preregistered a comprehensive set of hypotheses testing main effects and interactions across multiple experimental dimensions.

Using Images in Multifactorial Conjoint Experiments

Multifactorial survey experiments (or "conjoint" experiments) have been a boon for the social sciences (Auspurg and Hinz 2014). These designs simultaneously randomize many attributes, allowing researchers to efficiently estimate the causal effects of multiple factors within a single study while better approximating the complexity of real-world decision-making. By presenting respondents with profiles that vary across numerous dimensions, these designs force tradeoffs that reduce social desirability bias, as the complexity makes it difficult for respondents to identify socially acceptable responses. Additionally, conjoint experiments provide clear causal estimates through average marginal component effects (AMCEs), which

capture the marginal impact of changing one attribute while averaging over the distribution of all other attributes (Hainmueller, Hopkins, and Yamamoto 2014).

A review article by Bansak et al. (2021) points to one dimension for innovation in this method: "conjoint designs have been administered primarily via tables with written attribute values, even though information about political candidates or other choices is often processed through visual, auditory, or other modes" (p. 27).

We are inspired by recent work across the social sciences in the use of nontextual modalities to communicate part or all of the stimuli in factorial experiments. For example, our choice to operationalize the race/ethnicity and gender of the occupant with images is deliberate. Following the theoretical argument from Abrajano, Elmendorf, and Quinn (2018), we maintain that using textual labels to represent these demographic characteristics provides limited information and that respondents tend to fill in correlated information based on stereotypes; however, when gender and especially race/ethnicity are communicated with an image, the stimulus bundles further information that blunts the impact of these stereotypes.[3] Indeed, social science research has increasingly employed altered photographs as experimental tools to examine how individuals categorize and evaluate social actors. Valentino et al. (2019) systematically vary the complexion of hypothetical migrants in an 11-nation study to analyze how visible phenotypic differences influence responses to outgroup members. Building on this approach, Schachter, Flores, and Maghbouleh (2021) use comparable complexion modifications to isolate the distinct role of appearance in systems of racial classification, separating this factor from genealogical background and cultural signifiers. More recently, McClean and Ono (2024) edit images to modify the perceived age of public officials in conjoint designs, demonstrating how overlapping status characteristics shape collective assessments.

Advances in generative AI have opened up exciting possibilities for visual conjoint designs. López Ortega and Radojevic (2025) use this approach in combination with a standard candidate choice conjoint experiment. Methodologically, Schwitter (2025) provides valuable implementation details for systematically varying only the relevant portions of images used. We follow the advice to carefully assess whether visual stimuli genuinely enhance the research question—and indeed, we decided that the information about the occupant and their spouse's employment status was better operationalized with text.

The use of visual stimuli to represent something *other* than a person is quite rare; however, this is a major part of the methodological contribution by Thébaud et al. (2021). Vecchiato and Munger (2025) is an important exception from which we draw significant inspiration. They embed every facet of a conjoint experiment in a visual stimulus, in this case, a hypothetical candidate social media profile. Our central methodological innovation is to combine the visual representation of household tidiness with the logic of the visual conjoint and the technique of generative AI.

Methods

Experimental Design

Our design systematically manipulates visual and textual features across three information sources: the room, the occupant, and the employment status of the

This is a room in this person's home.

On a scale from 0 to 100, how messy is the room? Use the slider below to indicate your response.

The pictured person receives an unexpected social visit in this room. How do you think the room's condition may affect the visitor's opinion of the pictured person?

The pictured person lives in this home with his spouse. He is a homemaker, and she works full time.

Who do you think is primarily responsible for cleaning this room?

Figure 1: Full representation of the visual conjoint and survey flow.

occupant/spouse. For a visual representation of these sources of variation see Fig. 1. We retain the logic of the experimental design in Thébaud et al. (2021), streamlined and extended for our theoretical purposes, including the order of the outcome variables and the location of the different types of information (with employment information after the first set of dependent variables). Each vignette includes the two images, fully randomized across all possible conditions; respondents see these images and then answer the first two outcome questions about room messiness and social consequences. We then include the information about employment status, followed by the final outcome question about responsibility.

Each respondent evaluated five vignettes, with each vignette consisting of an AI-generated photorealistic image of a room, an AI-generated photorealistic image of the room's occupant, and text describing the occupant's and their (different-sex) spouse's employment status. As a result, we had to significantly reduce the number of outcome variables for each of the vignettes compared to Thébaud et al. (2021). We discuss these decisions later. Experimental factors were fully randomized both within and between respondents.

Visual Manipulations

Our visual stimuli manipulated eight factors across the room and occupant images.

Room characteristics: We varied four features of the room images:

- **Room tidiness**: messy versus tidy (two levels)
- **Room type**: kitchen versus living room (two levels)
- **Source of mess**: adult-caused mess versus child-caused mess (two levels)
- **SES**: higher SES versus lower SES, signaled through furnishing quality and style (two levels)

Occupant characteristics: We varied two features of the occupant images:

- **Gender presentation**: masculine versus feminine appearance (two levels)
- **Race/ethnicity**: white versus Black versus Hispanic appearance (three levels)

All room images were generated using Gemini Flash 2.5, with extensive prompt engineering to ensure consistent quality and appropriate variation within each experimental condition. Images were first reviewed by the authors and excluded if they failed to clearly represent the intended experimental manipulation. We generated multiple candidate images for each combination of room characteristics and selected those that best balanced realism with clear experimental contrast.

Similarly, occupant images were generated using Gemini Flash 2.5 with carefully constructed prompts to create photorealistic headshots that varied systematically in gender presentation and race/ethnicity while holding other factors as constant as possible. All occupant images depicted adults of roughly similar age wearing professional attire. See Appendix A for a full description of this process.

Text-Based Manipulations

After respondents answered initial questions about the room, we revealed text describing the employment status of both the pictured person and their spouse:

- **Occupant employment**: works full-time versus is a homemaker versus is unemployed (three levels)
- **Spouse employment**: works full-time versus is a homemaker versus is unemployed (three levels)

All vignettes described the occupant as living with a different-sex spouse, yielding nine possible employment combinations that were fully crossed with all other experimental factors—except that the pronouns describing the occupant were matched with the gender of the image. The text was presented in the following format: "The pictured person lives in this home with [his/her] spouse. [He/She] [works full-time / is a homemaker / is unemployed], and [his/her spouse] [works full-time / is a homemaker / is unemployed]."

Experimental Flow

For each vignette, respondents encountered the following sequence:

1. Room image and occupant image displayed with the prompt: "This is a room in this person's home."
2. Perceived messiness question: "On a scale from 0 to 100, how messy is the room?" (0 = not messy at all, 100 = extremely messy; continuous slider)[4]
3. Social consequences question: "The pictured person receives an unexpected social visit in this room. How do you think the room's condition may affect a visitor's opinion of the pictured person?" (5-point scale: makes the visitor's opinion much more negative; makes the visitor's opinion somewhat more negative; doesn't change the visitor's opinion; makes the visitor's opinion somewhat more positive; and makes the visitor's opinion much more positive)[5]
4. Employment status text revealed
5. Responsibility allocation question: "Who do you think is primarily responsible for cleaning this room?" (5-point scale: the pictured person; mostly the pictured person; equal responsibility; mostly the pictured person's spouse; and the pictured person's spouse)[6]

This staged revelation design allowed us to measure perceptions and anticipated social consequences before introducing employment information, which could prime respondents' judgments. The responsibility question was presented after employment information to capture how time availability and gender jointly shape attributions.

Sample and Recruitment

We recruited respondents through Prolific in November 2025. Prolific's quota sampling was used to approximate the U.S. population on gender, race/ethnicity, and age. Respondents were compensated at an average expected rate of $12 per hour, assuming the survey took four minutes to complete. The empirical mean duration was 3 minutes 49 seconds.

We implemented two standard text comprehension checks at the beginning of the survey. Respondents who failed either attention check were excluded before any treatment was implemented. We started with a sample of 3,078, of whom 52 were removed for failing an attention check and a further 32 did not complete the survey, yielding a final analytic sample of 2,994 respondents, who each viewed five vignettes for a total of 14,970 observations. This matches the expectation from our power analysis, which was conducted using the "cjpowR" package (Freitag and Schuessler 2020); we are powered to be able to detect interaction effects (not just main effects) with a standardized effect size of 0.05. Absent specific methodological guidance on this point, we increased the required sample size by 20 percent to account for the preregistered adaptive shrinkage adjustments for multiple comparisons.

Analytic Strategy

We estimate AMCEs using ordinary least squares regression with respondent-clustered standard errors (Hainmueller et al. 2014). For each outcome, we regress the outcome on indicator variables for each level of each experimental factor, excluding one reference category per factor. The AMCE represents the average change in the outcome when moving from the reference category to the specified level, averaged over all possible combinations of the other attributes.

For our primary analyses, we exclude the "room_state" (messy vs. tidy) factor from visualizations because its effect size is extremely large and would compress all other estimates. However, "room_state" remains in all regression models to ensure proper estimation of the remaining effects.

For the social consequences and responsibility outcomes, we convert the 5-point ordinal scales to numeric values (1–5) and analyze them as continuous outcomes, following standard practice in conjoint analysis (Bansak et al. 2021). For responsibility specifically, we create an alternative outcome that recodes responses to measure the responsibility attributed to women (rather than to the pictured person), reversing the scale when the pictured person is male. This allows us to directly estimate how gender affects responsibility attributions while accounting for the direction of the original question.

Where relevant, we test for variation in treatment effects by respondent characteristics by estimating marginal means (Leeper, Hobolt, and Tilley 2020), which are comparable across subgroups, unlike AMCEs. For the AMCE analyses, we preregistered that we would present both the uncorrected estimates and the Adaptive SHrinkage (ASH) estimator proposed by Liu and Shiraito (2023). This approach to multiple testing corrections aims to address the inflated risk of false positives inherent in testing many hypotheses simultaneously. This is a middle ground approach, less conservative than the Bonferroni but more conservative than the Benjamini–Hochberg, and more principled than either, adjusting not just the hypothesis testing procedure but also the point estimates.

All analyses were conducted in R version 4.4.2 using the packages `cjoint` (Strezhnev et al. 2018), `cregg` (Leeper 2020), and `tidyverse` (Wickham et al. 2019).

Results

Descriptive Overview

Figure 2 presents the distribution of perceived messiness ratings across our three treatment conditions that varied the source of mess: tidy rooms, messy rooms with child-caused mess, and messy rooms with adult-caused mess. As expected, tidy rooms received dramatically lower messiness ratings (mean = 2.86, SD = 12.1) compared to messy rooms, whether the mess was attributed to children (mean = 81.0, SD = 17.7) or adults (mean = 76.7, SD = 21.8). Child-caused mess was perceived as significantly messier than adult-caused mess. However, there was substantial heterogeneity within the messy room conditions. Some respondents rated even objectively messy rooms quite low on the messiness scale, whereas others maxed out the scale. This variation suggests that perceptions of messiness are not purely

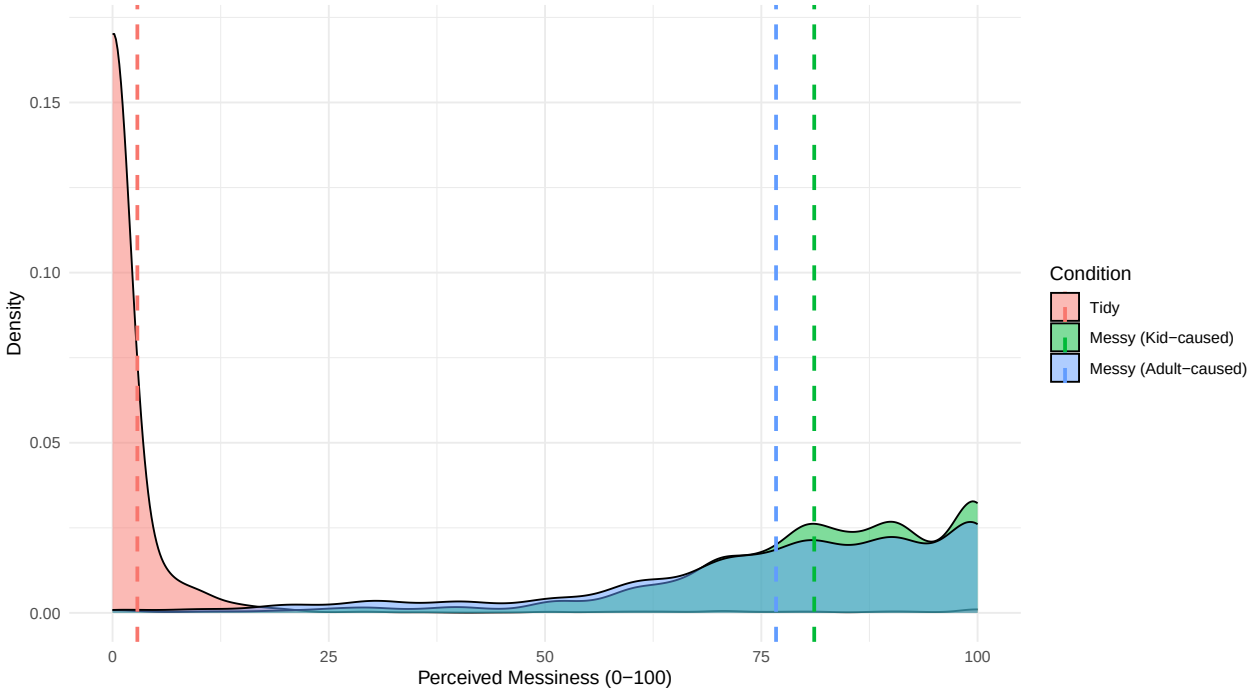

Figure 2: Density distribution of perceived messiness by treatment condition. Dashed vertical lines indicate mean values for each condition.

determined by objective features of the room, but also reflect respondent-level factors and potentially interactions with other experimental manipulations. In contrast, nearly 20 percent of respondents put tidy rooms at the minimum value and almost none of them ranked tidy rooms more than 20 out of 100, suggesting that they are not responding at random.

Main Effects on Perceived Messiness

Figure 3 presents AMCEs for all experimental factors on perceived messiness, excluding the visualization of the room tidiness factor, for the reasons discussed earlier. First, a child-caused mess is perceived as significantly messier than an adult-caused mess. Second, kitchens are judged as less messy than living rooms. Finally, the low-SES rooms are rated as less messy than the high-SES rooms; all three of these estimates remain statistically significant after the ASH adjustment (see Fig. 18).

Contrary to our expectations based on prior work (Thébaud et al. 2021), we find no significant main effect of occupant gender on perceived messiness. Women's rooms are rated similarly to men's rooms, averaging over the values of other variables. We return to this surprising null finding in the discussion. We find very small effects for occupant race/ethnicity. There is a marginally significant difference between white and Black occupants, with the former's rooms being rated as slightly messier, but this effect is not robust.[7] The employment status of the occupant and

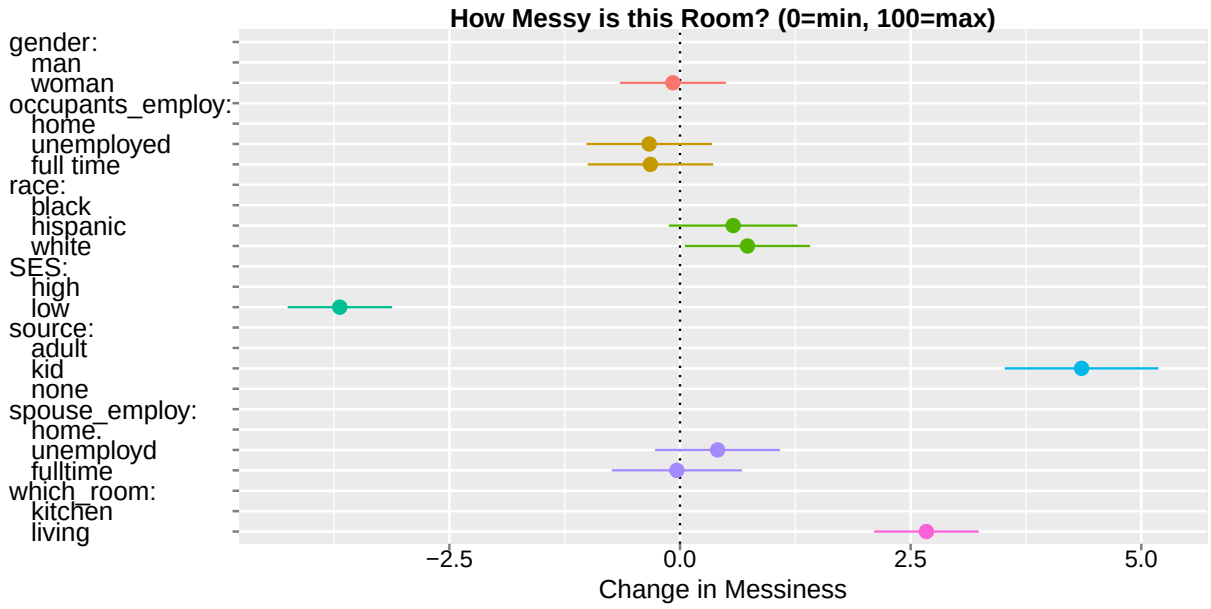

Figure 3: Average marginal component effects on perceived messiness (0–100 scale). Points represent point estimates with 95 percent confidence intervals. Reference categories are the first listed below the facet name. The room tidiness factor is excluded from visualization but included in the regression model. Standard errors clustered at the respondent level. N = 14,970.

their spouse each has no significant effect; this is encouraging given the survey flow, which asks respondents this question before they see these employment statuses.

Main Effects on Anticipated Social Consequences

Figure 4 presents effects on anticipated social consequences: how respondents expected a visitor to judge the room's occupant. Again, the AMCE for room status is extremely powerful (messy rooms carry a large social penalty compared to clean rooms) and thus not shown for legibility reasons. The pattern of results largely follows the logic of the room status results: the living room was perceived as messier and, therefore, causes greater social penalty and the low SES room was perceived as less messy and, therefore, causes a social bonus (though note that the magnitude of this effect is relatively smaller than the effect on perceptions of messiness, perhaps because SES has a direct effect on social perception independent of messiness).

The crucial exception to this logic is the source of mess: child-caused mess reduces negative social consequences compared to adult-caused mess, despite being perceived as messier. This suggests that child-caused mess operates as a legitimating excuse, mitigating social judgment even as it increases perceived disorder. The living room/kitchen and the child/adult results persist after the ASH correction, but the SES results become non-significant (Fig. 19).

As with perceived messiness, we find no main effect of occupant gender on social consequences. This null result is particularly surprising given Thébaud et al.

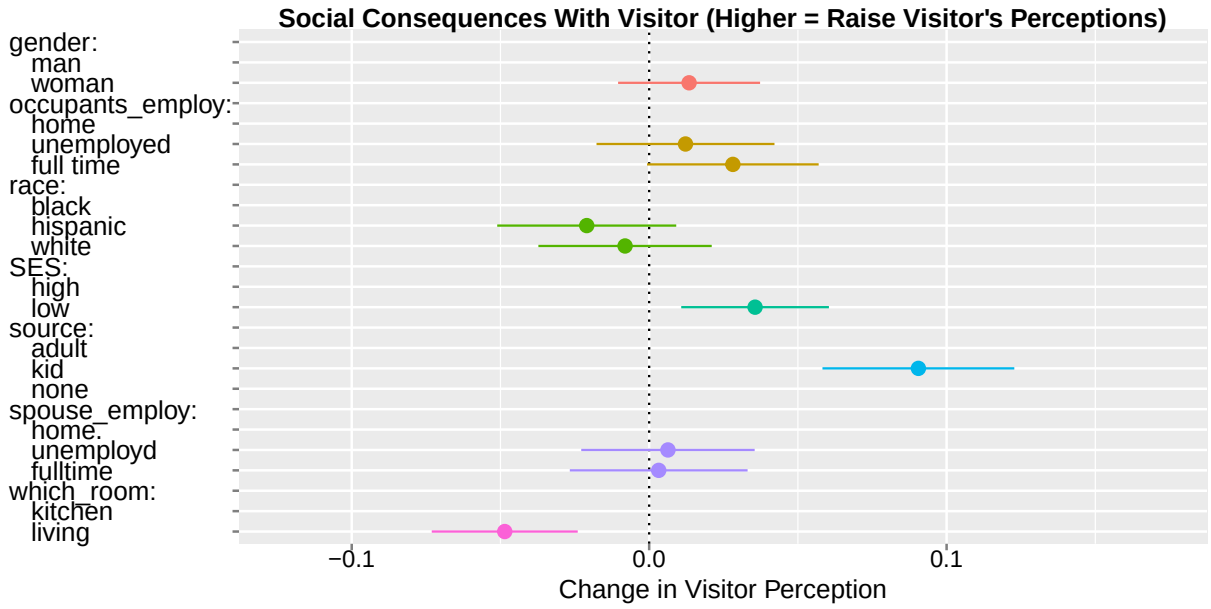

Figure 4: Average marginal component effects on anticipated social consequences (5-point scale from much more negative to much more positive visitor opinion). Points represent point estimates with 95 percent confidence intervals. Reference categories are the first listed below the facet name. The room tidiness factor is excluded from visualization but included in the regression model. Standard errors clustered at the respondent level. $N = 14{,}970$.

(2021)'s finding that women face harsher social judgment for mess. We explore this discrepancy further later.

Responsibility Attribution

Figure 5 presents effects on responsibility attribution. The largest effects are caused by employment status, with independent but symmetric effects for the employment status of the occupant and their spouse. Working full time is associated with an average of one full-point decrease in responsibility on the five-point scale. The unemployed see a small but statistically significant decrease in responsibility compared to homemakers, as we predicted. This difference between homemakers and the unemployed is only at the level of $p < 0.10$ and does not persist after the ASH correction, but the other significant results do (Fig. 20).

None of the features of the room or the race/ethnicity of the occupant has an effect, but here, we observe strong evidence for an effect of gender. Women are assigned significantly more cleaning responsibility than men, averaged across all values of employment status. The magnitude of this gender effect is roughly one-quarter of that of full-time employment.

To more directly estimate gender effects on responsibility while accounting for question framing, we recoded the outcome variable to measure the responsibility assigned by gender as well as the employment variables (in our always different-sex couples), regardless of occupant status. Figure 6 presents results using these

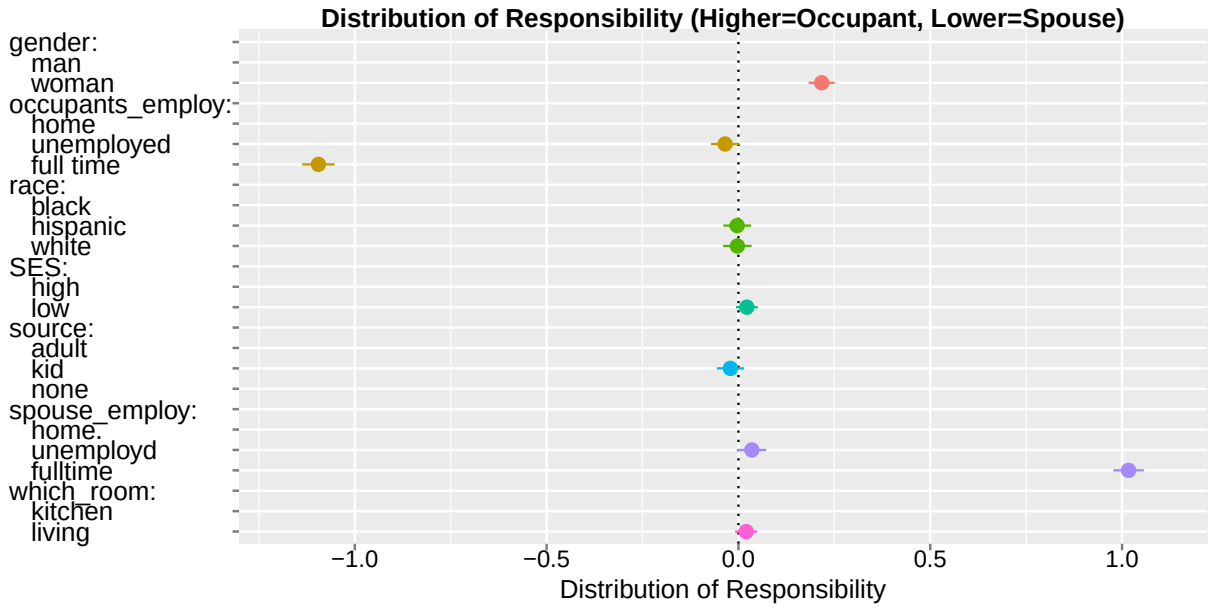

Figure 5: Average marginal component effects on responsibility attribution (5-point scale from pictured person's spouse to pictured person). Higher values indicate more responsibility assigned to the pictured person. Points represent point estimates with 95 percent confidence intervals. Reference categories are the first listed below the facet name. The room tidiness factor is excluded from visualization but included in the regression model. Standard errors clustered at the respondent level. N = 14,970.

recoded variables, along with the interaction between the man's employment and the woman's employment.

This specification reveals the employment interaction more clearly. The main effects are still symmetric by gender when we compare the "man employ" and "woman employ" facets. But the interaction term allows us to decompose the source of the gender penalty displayed in Figure 5.[8] When one person works full time, and the other is unemployed, there is no significant interaction effect (compared to the baseline where both are homemakers). However, when both are employed full time, the woman is assigned significantly more responsibility; this result persists after the ASH adjustment (Fig. 21). The magnitude of this interaction effect reveals that the large majority of the overall gender penalty comes from this combination.

This operationalization of the outcome also allows us to detect a small but significant effect of the source of the mess: when it is children's mess rather than adults' mess, women are assigned more responsibility for cleaning. This result also persists after the ASH adjustment.

Marginal Means Analysis For Respondent Gender

We note that the AMCE is a causal estimand resulting from the combination of treatments delivered by the visual conjoint; it is therefore inappropriate for purposes of understanding differences associated with respondent characteristics, which is a descriptive estimand (Lundberg, Johnson, and Stewart 2021). To examine

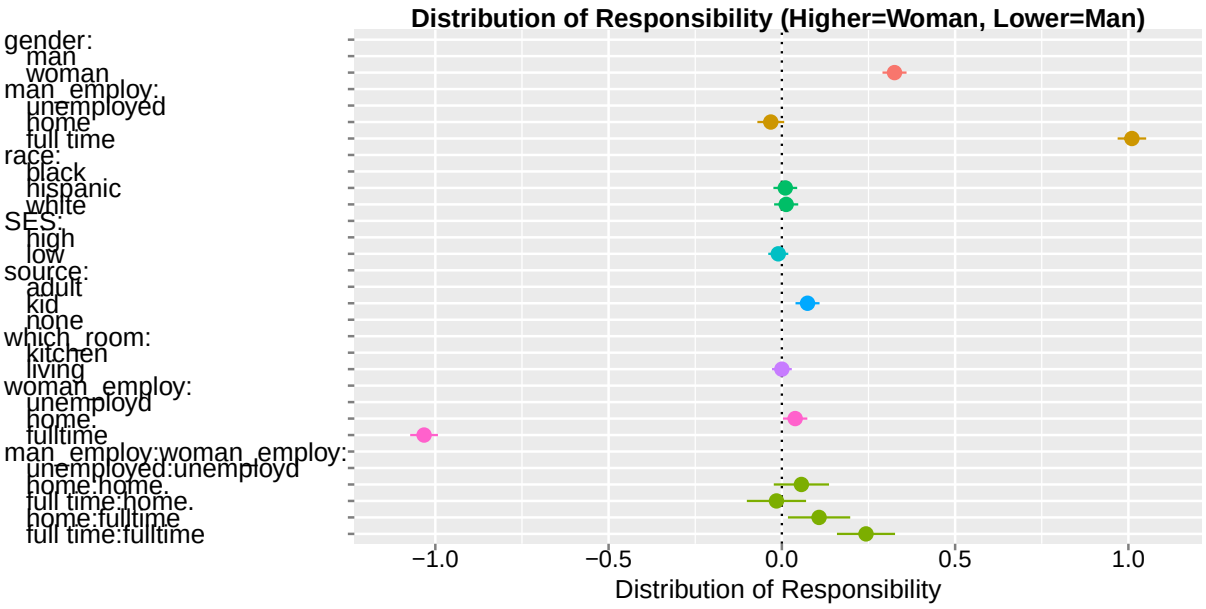

Figure 6: Average marginal component effects on responsibility attribution [5-point scale from men's responsibility (1) to woman's responsibility (5)]. Points represent point estimates with 95 percent confidence intervals. Reference categories are the first listed below the facet name. The room tidiness factor is excluded from visualization but included in the regression model. Standard errors clustered at the respondent level. N = 14,970.

differences by respondent sex, then, we conduct a marginal means analysis as suggested by Leeper et al. (2020). The results of this analysis are displayed in Figure 7. Note that our N decreases slightly, from 2,994 respondents to 2,982 respondents, when we remove the 12 respondents who selected "Prefer not to say" when asked for their gender.

We see that women are significantly more likely than men to think that cleaning this room is the woman's responsibility. This is a surprising result that we discuss further later. There are two other important results from this analysis. The first is the fact that the global average mean is 3.11, significantly above the midpoint. This represents part of the overall gender bias found in Figure 5. The other part comes from the marginal means estimated by the gender of the occupant in Figure 7. Theoretically, there should not be any differences associated with the gender of the occupant, as all are in different-sex couples; the observed differences are evidence that respondents are anchoring on the person pictured when making their evaluation of responsibility. This is a source of noise that represents a weak point in our design, though the overall average marginal mean and the interaction between the employment statuses of the couple (displayed in Fig. 6) demonstrate that we are still able to detect the hypothesized effects.

Marginal mean analysis to understand the association with respondent gender for the other two outcome variables can be found in Appendix B; we find no evidence that men and women respond differently on either room messiness or occupant social consequences.

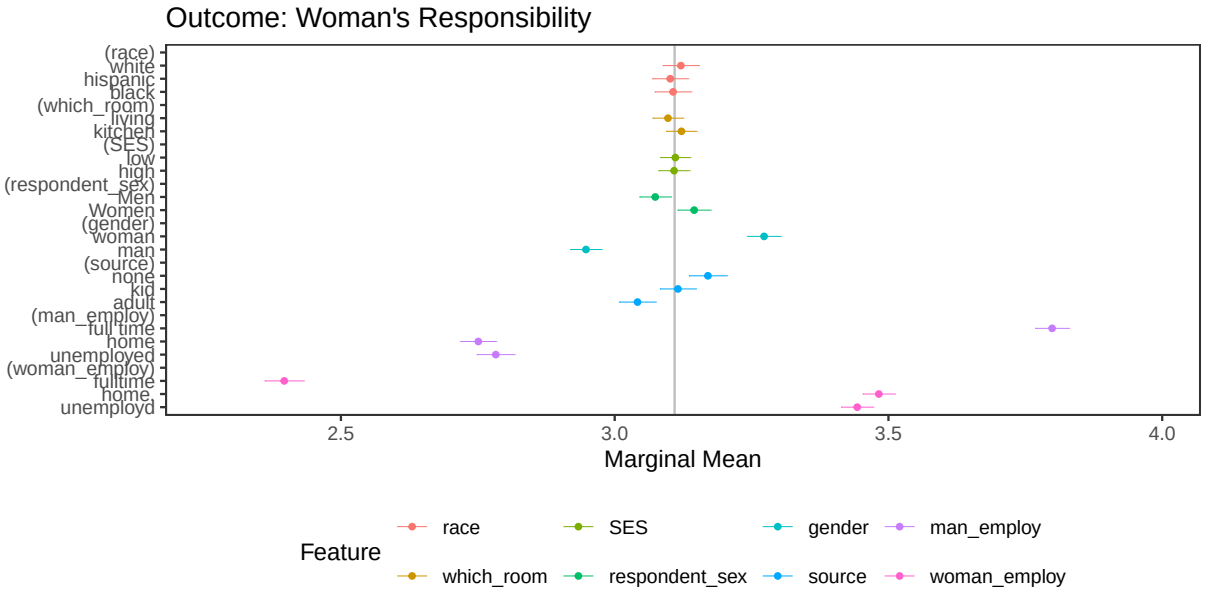

Figure 7: Marginal means for all factors predicting responsibility assigned to women. Marginal means represent the average predicted responsibility level for each factor level, marginalizing over all other factors. The vertical line indicates the overall mean. N = 14,910.

Interaction Effects

We preregistered the analysis of a number of interaction effects between the occupant gender and the other facets we experimentally manipulated. We find no evidence for any of these interaction effects, which are summarized in Table 1; full results can be found in Appendix C. One explanation for these null results is that the magnitude of the effect of the room status manipulation soaked up too much of the variation.

Conclusion

This study employed AI-generated visual conjoints to investigate the normative mechanisms underlying gender inequality in housework attribution. Building on

Table 1: Summary of Interaction Effects.

Treatment Dimension	Perceived Messiness	Social Consequences
Social identities		
Gender × race	Null	Null
Gender × SES	Null	Null
Spatial context		
Gender × room type	Null	Null

the theoretical framework and experimental logic of Thébaud et al. (2021), we extended prior work in several key dimensions: a substantially larger sample size, a conjoint design with multiple vignettes per respondent, visual operationalization of both room conditions and occupant characteristics, and novel experimental factors including room type, source of mess, SES, and the employment arrangements of both partners in a couple.

Our central finding is that gender effects on housework norms operate primarily through responsibility attribution rather than through perception or anticipated social judgment. Contrary to our expectations based on Thébaud et al. (2021), we find no evidence that women's rooms are perceived as messier than men's rooms, nor that women face harsher anticipated social consequences for identical levels of domestic disorder. However, we find robust evidence that women are assigned significantly more cleaning responsibility than men, even after accounting for employment status. This gender penalty is particularly pronounced when both partners work full time—precisely the configuration in which egalitarian norms might be expected to prevail.

How should we reconcile our null findings on perception and social judgment with the significant effects documented by Thébaud et al. (2021)? We caution that we can only offer our interpretation of these discrepancies rather than empirical evidence, and we encourage further fine-grained experimental work to test these interpretations. We believe that the explanation is primarily methodological rather than substantive; several design differences may account for the divergent findings. First, Thébaud et al. (2021) operationalized occupant gender through first names, whereas we used AI-generated photorealistic images. Following the theoretical argument in Abrajano et al. (2018), visual stimuli bundle additional information that may block the impact of stereotypes activated by textual labels alone. Moreover, first names themselves carry signals beyond gender, including information about SES and race/ethnicity (Crabtree et al. 2022; Gaddis 2017). Second, our conjoint design presented respondents with multiple vignettes varying across many dimensions simultaneously, which may have reduced the salience of any single attribute—including gender—on perception and judgment outcomes. The complexity of the conjoint task may have attenuated effects that emerge more clearly in simpler experimental designs.

Importantly, the fact that we find strong gender effects on responsibility attribution despite null effects on perception and judgment suggests that these normative mechanisms may be partially independent. Respondents in our study did not perceive women's rooms as messier or anticipate that women would face harsher judgment; yet, they nonetheless assigned women greater responsibility for cleaning. This pattern implies that responsibility attribution may operate through a distinct cognitive pathway-one that directly invokes gender role expectations without necessarily passing through differential perception or anticipated sanction. Future research might profitably investigate the conditions under which these mechanisms align versus diverge.

Beyond the gender effects, our study yielded several novel findings. Child-caused mess presents an intriguing case: it is perceived as significantly messier than adult-caused mess; yet, it reduces anticipated negative social consequences.

This suggests that children's mess operates as a legitimating excuse, mitigating social judgment even as it increases perceived disorder. The presence of children may signal situational constraints that excuse domestic disorder, consistent with accounts emphasizing the role of attributions in social judgment. We find no evidence that child-caused mess differentially affects judgments of male versus female occupants. Notably, however, we do find that women are held more responsible for child-caused mess. This effect is small and was not preregistered, so it should be interpreted with caution. Nevertheless, it does align with theoretical frameworks that emphasize the intersection of gender and parental status in domestic responsibility norms.

Employment status emerged as the strongest predictor of responsibility attribution, with symmetric effects for occupants and their spouses: working full time substantially reduces attributed responsibility, whereas homemaker status increases it (though this latter finding is only marginally significant and does not hold when we perform the ASH adjustment). This pattern is consistent with time availability explanations for housework division. However, the interaction between partners' employment statuses reveals a more nuanced picture. When both partners work full time, women are assigned significantly more responsibility than the symmetric main effects would predict. This accounts for the majority of the overall gender penalty we observe. This finding provides empirical support for theoretical and descriptive accounts of the "second shift" (Hochschild and Machung 1989) and compensatory gender display (Brines 1994; Tichenor 2005). When women's employment matches or exceeds their male partners', the persistence of gendered domestic responsibility suggests that egalitarian work arrangements do not automatically translate into egalitarian household arrangements—indeed, gender norms may reassert themselves precisely when women's breadwinning threatens traditional gender hierarchies.

Our marginal means analysis by respondent gender yielded a surprising result: women respondents assign significantly more cleaning responsibility to women than men respondents do. One interpretation of this finding draws on an internalized norms account. In our theoretical framework, we operationalized the socialization mechanism using respondents' evaluations of messiness to capture the idea that women and men might see mess differently. But, in fact, responsibility assignment might be another way to tap into the same theoretical mechanism. If women have been socialized to view domestic maintenance as a feminine responsibility, they may apply these standards when evaluating others. We note that this finding was not preregistered and should be interpreted as tentative and as a productive avenue for future investigation.

We found little evidence for effects of occupant race/ethnicity or SES on our primary outcomes. The theoretical predictions for these factors were admittedly ambiguous—competing mechanisms predicted effects in opposite directions—and our null findings do not resolve these theoretical debates. It is possible that our visual operationalizations of race/ethnicity and SES were insufficiently strong or salient to activate the relevant stereotypes and expectations. Alternatively, these factors may genuinely exert weaker effects on housework-related judgments than gender and employment status, at least in the experimental context we created. We

also note that the large effect of room tidiness may have absorbed variance that would otherwise reveal more subtle effects of occupant characteristics.

Methodologically, this study demonstrates the promise of AI-generated visual stimuli for conjoint experiments. Generative AI allowed us to create photorealistic images that systematically varied multiple dimensions while holding irrelevant factors constant—a task that would be extremely difficult with traditional photography. The iterative process of prompt engineering, combined with validation by multiple AI models, enabled us to achieve reasonable experimental control over complex visual stimuli. We echo the recommendations of Schwitter (2025) regarding careful assessment of whether visual stimuli genuinely enhance the research question, and we note that our decision to operationalize employment status through text rather than images reflects a judgment that some factors are better communicated through one modality than another. As generative AI capabilities continue to advance, we anticipate growing opportunities for visual conjoint designs across the social sciences.

However, limitations in our understanding of the methodological functioning of these AI-generated images as stimuli warrant acknowledgment. Despite our efforts to maximize the realism of the images, it is possible that some respondents suspected that the images were AI-generated rather than photographs and that this changed how they responded to the stimuli. Although we did not test this point explicitly, we note some recent experimental evidence showing that AI-generated media have broadly the same effect (in this case, on the legitimacy of targeted politicians) even among respondents who identify them as AI-generated (Hameleers et al. 2026). We can also report that none of our respondents used the Prolific comment section to raise concerns about AI-generated stimuli, though several of them had comments about other aspects of the experiment.

From a larger methodological perspective, we note the tension between innovation and validation in experimental design. We intentionally kept much of our design similar to that of Thébaud et al. (2021) to enhance the comparability of our results. However, we innovate in both the theoretical content and the technology of stimulus delivery, limiting our ability to directly test for differences between photographic and AI-generated images. As AI-generated images become taken up by survey experimentalists, we expect to see further, detailed validation work to understand how survey participants understand this kind of stimulus.

Despite these limitations, our findings advance understanding of the normative mechanisms underlying persistent gender inequality in housework. The fact that gender effects emerge in responsibility attribution but not in perception or anticipated judgment reveals the specific pathway through which gender norms operate. Our results point to both structural and cultural levers for change. First, the powerful effects of time availability through different levels of employment suggest that policies supporting women's labor force participation can meaningfully reduce their housework burden by establishing legitimate claims on time. However, our finding that women are assigned greater responsibility precisely when both partners work full time reveals the limits of structural solutions alone. Even when time availability is equalized, deeply held cultural expectations about domestic responsibility persist. As progress toward gender equality in the domestic sphere continues to lag behind convergence in education and employment, understanding

both the structural and cultural mechanisms through which inequality persists remains essential for identifying effective paths forward.

Data Availability

Anonymized data and full replication code are available on the author's github https://github.com/kmunger/Housekeeping_SocSci_Replication.

Ethical Statement

This experiment was approved by the Ethics Committee at the European University Institute, reference code 20250430.

AI Usage Statement

In addition to the use of AI to generate and validate the visual stimuli (detailed in the Online Appendix), we used Claude Opus 4.5 and 4.6 to write the Javascript for the custom Qualtrics file, proofread the manuscript, interface with our bibtex file for citation management, write a first draft of the abstract, and clean up and compile the R code for the Replication Materials.

Notes

1 We cluster our standard errors at the respondent level, so our effective sample size is somewhat smaller than this multiple.

2 Available at: https://osf.io/7k8e6/.

3 This argument accords with evidence about the role of the amount of information provided in factorial survey experiments (Auspurg, Hinz, and Sauer 2017a).

4 Thébaud et al. (2021) include a second question here, about how urgent it would be for the occupant to clean the room. Their results for the two questions are broadly parallel, and we found little theoretical difference between these two outcomes, so we ask only the household tidiness question.

5 Thébaud et al. (2021) ask about three specific types of visitors; the reliability coefficient for these questions is very high and so we simply ask about a hypothetical unspecified visitor. They also ask a version of this question, which asks the respondent to predict the hypothetical occupant's level of comfort with a hypothetical guest; as mentioned earlier, the results are parallel for these two question wordings, and we prefer the question about actual social consequences for the sake of parsimony. They also ask a battery of questions about the moral character of the occupant. We were concerned that this outcome was too far removed from the experimental stimuli, as evidenced by the fact that the results in the original paper are significant in an unexpected direction: female occupants of messy rooms are rated as more moral than male occupants of messy rooms, but female occupants of clean rooms are rated as less moral than male occupants of clean rooms. In light of this, we dropped the questions about morality.

6 Thébaud et al. (2021) varies whether the occupant is single or partnered, if they have children or not, and whether they work full time or not. We provide information about

parenthood status via the image of the room, and we focus only on partnered occupants. This allows us to systematically vary both the employment status of the occupant *and* their partner, which is essential for our theoretical innovation.

7 See the marginal means analysis in Appendix B. This approach finds no significant effects for any of the racial/ethnic categories, and the original finding is not robust to the ASH adjustment in Appendix D.

8 Note that because of collinearity, we are only able to estimate four of the eight possible interacted contrasts. Setting the baseline category where both people are unemployed, we can estimate four of the interacted contrasts where neither person is unemployed.

References

Abrajano, Marisa A., Christopher S. Elmendorf, and Kevin M. Quinn. 2018. "Labels vs. Pictures: Treatment-Mode Effects in Experiments about Discrimination." *Political Analysis*, 26:20–33. https://doi.org/10.1017/pan.2017.36

Altintas, Evrim and Oriel Sullivan. 2016. "Fifty Years of Change Updated: Cross-National Gender Convergence in Housework." *Demographic Research* 35:455–70. https://doi.org/10.4054/DemRes.2016.35.16

Auspurg, Katrin and Sabine Düval. 2024. "Housework as a Woman's Job? What Looks Like Gender Ideologies Could Also Be Stereotypes." *Sociological Science* 11:789–814. https://doi.org/10.15195/v11.a29

Auspurg, Katrin and Thomas Hinz. 2014. *Factorial Survey Experiments*. Vol. 175. Sage Publications. https://doi.org/10.4135/9781483398075

Auspurg, Katrin, Thomas Hinz, and Carsten Sauer. 2017a. "Why Should Women Get Less? Evidence on the Gender Pay Gap from Multifactorial Survey Experiments." *American Sociological Review* 82:179–210. https://doi.org/10.1177/0003122416683393

Auspurg, Katrin, Maria Iacovou, and Cheti Nicoletti. 2017b. "Housework Share Between Partners: Experimental Evidence on Gender-Specific Preferences." *Social Science Research* 66:118–39. https://doi.org/10.1016/j.ssresearch.2017.01.003

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

Berk, Sarah F. 1985. *The Gender Factory: The Apportionment of Work in American Households.* Springer Science & Business Media. https://doi.org/10.1007/978-1-4613-2393-8

Bittman, Michael, Paula England, Liana Sayer, Nancy Folbre, and George Matheson. 2003. "When Does Gender Trump Money? Bargaining and Time in Household Work." *American Journal of Sociology* 109:186–214. https://doi.org/10.1086/378341

Blau, Francine D., Lawrence Kahn, Matthew L. Comey, Amanda R. Eng, Pamela A. Meyerhofer, and Alexander Willén. 2020. "Culture and Gender Allocation of Tasks: Source Country Characteristics and the Division of Non-Market Work among US Immigrants." *Review of Economics of the Household* 18:907–58. https://doi.org/10.3386/w26931

Brines, Julie. 1994. "Economic Dependency, Gender, and the Division of Labor at Home." *American Journal of Sociology* 100:652–88. https://doi.org/10.1086/230577

Carriero, Renzo and Lorenzo Todesco. 2017. "The Interplay between Equity and Gender Ideology in Perceived Housework Fairness: Evidence from an Experimental Vignette Design." *Sociological Inquiry* 87:561–85. https://doi.org/10.1111/soin.12162

Crabtree, Charles, S. Michael Gaddis, John B. Holbein, and Edvard N. Larsen. 2022. "Racially Distinctive Names Signal Both Race/Ethnicity and Social Class." *Sociological Science* 9:454–72. https://doi.org/10.15195/v9.a18

Cunningham, Mick. 2005. "Gender in Cohabitation and Marriage: The Influence of Gender Ideology on Housework Allocation over the Life Course." *Journal of Family Issues* 26:1037–61. https://doi.org/10.1177/0192513X04273592

Doan, Long and Natasha Quadlin. 2019. "Partner Characteristics and Perceptions of Responsibility for Housework and Child Care." *Journal of Marriage and Family* 81:145–63. https://doi.org/10.1111/jomf.12526

Doan, Long, Natasha Quadlin, and Katharine Khanna. 2024. "Using Experiments to Study Families and Intimate Relationships," *Journal of Marriage and Family* 86:1251–71. https://doi.org/10.1111/jomf.12959

Dominguez-Folgueras, Marta. 2022. "It's about Gender: A Critical Review of the Literature on the Domestic Division of Work," *Journal of Family Theory & Review* 14:79–96. https://doi.org/10.1111/jftr.12447

England, Paula, Andrew Levine, and Emma Mishel. 2020. "Progress toward Gender Equality in the United States Has Slowed or Stalled." *Proceedings of the National Academy of Sciences* 117:6990–97. https://doi.org/10.1073/pnas.1918891117

Ferree, Myra M. 2010. "Filling the Glass: Gender Perspectives on Families." *Journal of Marriage and Family* 72:420–39. https://doi.org/10.1111/j.1741-3737.2010.00711.x

Few-Demo, April L. and Katherine R. Allen. 2020. "Gender, Feminist, and Intersectional Perspectives on Families: A Decade in Review," *Journal of Marriage and Family* 82:326–45. https://doi.org/10.1111/jomf.12638

Freitag, M. and Schuessler, J. (2020), "cjpowr-a priori power analyses for conjoint experiments," *R Package*. https://doi.org/10.31235/osf.io/9yuhp

Gaddis, S. Michael. 2017. "How Black Are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies." *Sociological Science* 4:469–89. https://doi.org/10.15195/v4.a19

Goldin, Claudia. 2014. "A Grand Gender Convergence: Its Last Chapter." *American Economic Review* 104:1091–19. https://doi.org/10.1257/aer.104.4.1091

Greenstein, Theodore N. 2000. "Economic Dependence, Gender, and the Division of Labor in the Home: A Replication and Extension." *Journal of Marriage and Family* 62:322–35. https://doi.org/10.1111/j.1741-3737.2000.00322.x

Hainmueller, Jens, Daniel J. Hopkins, and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22:1–30. https://doi.org/10.1093/pan/mpt024

Hameleers, Michael, Toni G. L. A. Van Der Meer, Marina Tulin, and Tom Dobber. 2026. "Radical Right-Wing Political Deepfakes Can Successfully Delegitimize Targeted Political Actors: Evidence From Three-wave Experiments in the US and The Netherlands." *Communication Research* 00936502261421437. https://doi.org/10.1177/00936502261421437

Hochschild, Arlie R. and Anne Machung. 1989. *The Second Shift: Working Parents and the Revolution at Home.* New York: Viking.

Hook, Jennifer L. 2017. "Women's Housework: New Tests of Time and Money," *Journal of Marriage and Family* 79:179–98. https://doi.org/10.1111/jomf.12351

Leeper, Thomas J. 2020. *cregg: Simple Conjoint Analyses and Visualization*, r package.

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

Liu, Guoer and Yuki Shiraito. 2023. "Multiple Hypothesis Testing in Conjoint Analysis." *Political Analysis* 31:380–95. https://doi.org/10.1017/pan.2022.30

López Ortega, Alberto and Marco Radojevic. 2025. "Visual Conjoint vs. Text Conjoint and the Differential Discriminatory Effect of (Visible) Social Categories." *Political Behavior* 47:335–53. https://doi.org/10.1007/s11109-024-09953-7

Lundberg, Ian, Rebecca Johnson, and Brandon M. Stewart. 2021. "What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory." *American Sociological Review* 86:532–65. https://doi.org/10.1177/00031224211004187

McCall, Leslie. 2005. "The Complexity of Intersectionality." *Signs: Journal of Women in Culture and Society* 30:1771–1800. https://doi.org/10.1086/426800

McClean, Charles T. and Yoshikuni Ono. 2024. "Too Young to Run? Voter Evaluations of the Age of Candidates." *Political Behavior* 46:2333–55. https://doi.org/10.1007/s11109-024-09920-2

Milkie, Melissa A., Liana C. Sayer, Kei Nomaguchi, and Hope X. Yan. 2025. "Who's Doing the Housework and Childcare in America Now? Differential Convergence in Twenty-First-Century Gender Gaps in Home Tasks." *Socius* 11:23780231251314667. https://doi.org/10.1177/23780231251314667

Pedulla, David S. and Sarah Thébaud. 2015. "Can We Finish the Revolution? Gender, Work-Family Ideals, and Institutional Constraint." *American Sociological Review* 80:116–39. https://doi.org/10.1177/0003122414564008

Pepin, Joanna R. and William J. Scarborough. 2025. "What Makes a Decision Fair? Relative Earnings, Gender, and Justifications for Couples' Decision-Making." *American Journal of Sociology* 130:1435–76. https://doi.org/10.1086/735618

Pessin, Léa. 2024. "Gender Equality for Whom? The Changing College Education Gradients of the Division of Paid Work and Housework among US Couples, 1968–2019." *Social Forces* 103:129–52. https://doi.org/10.1093/sf/soae028

Pessin, Léa and Bruno Arpino. 2018. "Navigating between Two Cultures: Immigrants' Gender Attitudes toward Working Women." *Demographic Research* 38:967. https://doi.org/10.4054/DemRes.2018.38.35

Pessin, Léa and Elena M. Pojman. 2025. "Racial-Ethnic Stratification in Work–Family Arrangements among Black, Hispanic, and White Couples." *Journal of Marriage and Family* 87:322–45. https://doi.org/10.1111/jomf.13020

Ridgeway, Cecilia L. 2011. *Framed by Gender: How Gender Inequality Persists in the Modern World*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199755776.001.0001

Ridgeway, Cecilia L. and Shelley J. Correll. 2004. "Unpacking the Gender System: A Theoretical Perspective on Gender Beliefs and Social Relations." *Gender & Society* 18:510–31. https://doi.org/10.1177/0891243204265269

Risman, Barbara J. 1998. *Gender Vertigo: American Families in Transition*. Yale University Press.

Risman, Barbara J. 2004. "Gender As a Social Structure: Theory Wrestling with Activism." *Gender & Society* 18:429–50. https://doi.org/10.1177/0891243204265349

Ross-Plourde, Mylène and Mylène Lachance-Grzela. 2025. "The Gendered Division of Housework in North America: A Systematic Review from 2014 to 2024." *Journal of Family Theory & Review*. https://doi.org/10.1111/jftr.70000

Schachter, Ariela, René Flores, and Neda Maghbouleh. 2021. "Ancestry, Color, or Culture? How Whites Racially Classify Others in the US," *American Journal of Sociology* 126:1220–63. https://doi.org/10.1086/714215

Schulz, Florian. 2021. "Attitudes towards Sharing Housework in Context: An Empirical, Factorial Survey Approach." *Journal of Family Research (JFR)* 33. https://doi.org/10.20377/jfr-419

Schwitter, Nicole. 2025. "Using Artificial Intelligence to Generate Visual Vignettes in Factorial Survey Experiments." *Social Science Computer Review* 08944393251392916. https://doi.org/10.1177/08944393251392916

Strezhnev, Anton, Jens Hainmueller, Daniel Hopkins, and Teppei Yamamoto. 2018. *cjoint: AMCE Estimator for Conjoint Experiments*. r package version 2.1.1.

Sullivan, Oriel. 2021. "The Gender Division of Housework and Child Care." Pp. 342–54 in *Research Handbook on the Sociology of the Family*. Edward Elgar Publishing. https://doi.org/10.4337/9781788975544.00033

Thébaud, Sarah, Sabino Kornrich, and Leah Ruppanner. 2021. "Good Housekeeping, Great Expectations: Gender and Housework Norms." *Sociological Methods & Research* 50:1186–214. https://doi.org/10.1177/0049124119852395

Tichenor, Veronica J. 2005. "Maintaining Men's Dominance: Negotiating Identity and Power When She Earns More." *Sex Roles* 53:191–205. https://doi.org/10.1007/s11199-005-5678-2

Urbina, Daniela R. and Daria Tisch. 2025. "Perceived Fairness of Couples' Division of Housework: Evidence From a Multi-Factorial Experiment in the United States." *Gender & Society* 08912432251337416. https://doi.org/10.1177/08912432251337416

Valentino, Nicholas A., Stuart Soroka, Shanto Iyengar, Toril Aalberg, Raymond Duch, Marta Fraile, Kyu S. Hahn, Kasper M. Hansen, Allison Harell, Marc Helbling, et al. 2019. "Economic and Cultural Drivers of Immigrant Support Worldwide." *British Journal of Political Science* 49:1201–26. https://doi.org/10.1017/S000712341700031X

Van Breeschoten, Leonie, Anne Roeters, and Tanja van der Lippe. 2018. "Reasons to Reduce: A Vignette-Experiment Examining Men and Women's Considerations to Scale Back following Childbirth." *Social Politics: International Studies in Gender, State & Society* 25:169–200. https://doi.org/10.1093/sp/jxy003

Vecchiato, Alessandro and Kevin Munger. 2025. "Introducing the Visual Conjoint, with an Application to Candidate Evaluation on Social Media." *Journal of Experimental Political Science* 12:57–71. https://doi.org/10.1017/XPS.2024.15

West, Candace and Don H. Zimmerman. 1987. "Doing Gender," *Gender & Society* 1:125–51. https://doi.org/10.1177/0891243287001002002

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D. McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas L. Pedersen, Evan Miller, Stefan M. Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana P. Seidel, Vitalie Spinu, Kohske Takahashi, David Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2019. "Welcome to the Tidyverse." *Journal of Open Source Software* 4:1686. https://doi.org/10.21105/joss.01686

Acknowledgments: This work has benefited from generous feedback from the Bocconi Dondena Seminar Speaker Series and the LMU Munich Department of Sociology Research Colloquium. Pessin acknowledges funding by the European Union, under the European Research Council grant for the WeEqualize project (grant agreement no. 101117327). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Léa Pessin: Department of Social and Political Science, European University Institute, Fiesole, Italy. E-mail: lea.pessin@eui.eu.

Kevin Munger: Department of Social and Political Science, European University Institute, Fiesole, Italy. E-mail: kevin.munger@eui.eu