Title

Trust in the wild: Multi-modal safety perceptions in naturalistic AV deployment

Journal

Transportation Research Part F: Traffic Psychology and Behaviour, 122

Authors

Guirado, Carlos
Oum, SangHyouk
Griswold, Julia
et al.

Publication Date

2026-08-01

Peer reviewed

Transportation Research Part F: Psychology and Behaviour 122 (2026) 103738

Contents lists available at ScienceDirect

Transportation Research Part F: Psychology and Behaviour

journal homepage: www.elsevier.com/locate/trf

Trust in the wild: Multi-modal safety perceptions in naturalistic AV deployment

Carlos Guirado [a,b,*], SangHyouk Oum [a], Julia Griswold [a], Joan Walker [b]

[a] Safe Transportation Research and Education Center (SafeTREC), University of California, Berkeley, United States
[b] Department of Civil and Environmental Engineering, University of California, Berkeley, United States

ARTICLE INFO

Keywords:
Autonomous vehicles
Safety perception
Trust in automation
Experience effects
Human-AV interaction
Driver/pedestrian psychology

ABSTRACT

Public acceptance remains a major barrier to autonomous vehicle (AV) deployment, yet most perception research relies on hypothetical surveys or small controlled experiments with AV-naive respondents. This cross-sectional study surveys 811 San Francisco Bay Area residents with real-world AV exposure (376 robotaxi passengers and 435 non-passengers: pedestrians, cyclists, and drivers) in one of the world's most mature deployment environments. We ask whether trust levels among AV-exposed residents differ from national baselines, whether trust varies by interaction mode, and whether experience is associated with differential recalibration of specific safety concerns. We find that 61% trust AVs to be safe, 4.7 times the 13% national rate. Trust varies sharply by mode: passengers report 82% trust versus 43-45% among non-passengers. Even non-riders far exceed national levels, consistent with ambient exposure beyond ridership alone contributing to higher trust, though residential selection cannot be ruled out. Concern about system malfunction is significantly lower among passengers (45%) than pedestrians and cyclists (63%), consistent with recalibration through direct performance feedback, whereas concern about unpredictable human behavior remains constant across all modes ($\sim 55\%$). These patterns suggest that experience is associated with attenuation of technology-related concerns while human-behavior concerns persist: a divergence with implications for communication strategies and infrastructure design as AV deployment expands.

1. Introduction

Autonomous vehicles (AVs) promise a fundamental transformation in road safety. Human error contributes to an estimated 94% of crashes (NHTSA, 2008), and AVs (which do not become intoxicated, distracted, or fatigued) could prevent tens of thousands of fatalities annually (Fagnant & Kockelman, 2015). Yet national surveys consistently show that most Americans fear self-driving vehicles: only 13% trust AVs to be safe, while 60% report being afraid of them (AAA Foundation, 2025). These figures have remained stable for years (AAA Foundation, 2023, 2024), raising a fundamental question: why does public distrust persist despite continued technological maturation and millions of miles of real-world operation?

One explanation is methodological: much of the existing AV perception research captures attitudes from respondents who have never encountered an AV. Pre-exposure survey results often rest on abstraction rather than direct evidence, yet behavioral psychology has long documented disconnects between stated intentions and actual behavior regarding novel technologies

* Corresponding author.
E-mail address: guirado@berkeley.edu (C. Guirado).

https://doi.org/10.1016/j.trf.2026.103738
Received 21 December 2025; Received in revised form 28 May 2026; Accepted 12 July 2026
Available online 30 July 2026
1369-8478/

(Conner & Norman, 2022; Sheeran, 2002). Indeed, when AV-experienced respondents are studied, substantial discrepancies emerge between their stated pre-exposure distrust and their actual behavior or trust levels (Fink et al., 2025; Power, 2024).

Importantly, real-world exposure to AVs need not require riding. Residents of cities with sustained AV deployment interact with these vehicles daily, watching them negotiate intersections, yield to pedestrians, and merge into traffic. This *ambient exposure* may itself shape attitudes (Zajonc, 1968), even among those who have never been passengers. If so, the relationship between experience and trust may thus be more nuanced than a binary of having ridden versus not.

Furthermore, trust is not monolithic. Lee and See's (2004) trust calibration framework, extended by Hoff and Bashir (2015), suggests that exposure might selectively recalibrate specific concerns rather than uniformly raising or lowering global trust. For example, while riding in an AV might alleviate fears about technology failure through direct performance feedback, concerns about human unpredictability in mixed traffic might remain intact because they lie outside the vehicle's performance loop (Kim & Kelley-Baker, 2021; Kyriakidis et al., 2015; Mao et al., 2025).

San Francisco offers a unique natural laboratory to examine these questions. Commercial robotaxi service launched in 2021 and expanded rapidly; by our study period (May-June 2025), Waymo operated as the sole commercial provider with more than 600 vehicles and citywide authorization, conducting approximately 150,000 paid trips per week (Commission, 2025; Examiner, 2025). Residents encounter AVs routinely: as passengers using commercial services, pedestrians and cyclists sharing crosswalks, or drivers sharing roads.

Our study addresses three gaps in the existing literature. First, we document perceptions among people who interact with AVs in daily life, moving beyond the hypothetical and controlled scenarios that dominate prior work (Gkartzonikas & Gkritza, 2019). Second, we expand the focus from passengers to the entire road-user ecosystem. Third, we examine whether experience differentially recalibrates specific concerns about technology failure versus human behavior. To our knowledge, this study provides the first systematic empirical documentation of AV safety perceptions across interaction modes in a naturalistic, saturation-level commercial deployment. Rather than proposing a new theoretical model, we extend empirical evidence within existing theoretical frameworks of trust calibration. This study addresses the following research questions:

1. Do trust levels among AV-exposed residents exceed national baselines, and do non-riders also show elevated trust?
2. Does trust vary by interaction mode (passenger, pedestrian, cyclist, driver)?
3. Among passengers, does cumulative experience predict higher trust? Among non-passengers, are close calls associated with lower trust?
4. Do concern patterns differ by mode? Specifically, do technology-related concerns attenuate with passenger experience, and do human-behavior concerns persist?

These questions examine how trust patterns form in a mature deployment setting where AVs are part of daily life, not why distrust persists in their absence. Thus, this study addresses one mechanism behind public distrust, not the full national acceptance puzzle.

The remainder of this paper is structured as follows. Section 2 reviews the literature on AV perceptions, trust calibration, and multi-modal road user interactions. Section 3 describes our survey methodology and approach. Section 4 presents results, and Section 5 discusses implications and limitations.

2. Literature review

This review synthesizes evidence on AV safety perceptions across five areas: the gap between hypothetical and experiential attitudes (Section 2.1), trust calibration theory and the challenge of separating experience effects from selection (Section 2.2), the structure of safety concerns (Section 2.3), the understudied perspective of non-passengers (Section 2.4), and theoretical predictions of behavior in mixed AV-human traffic (Section 2.5). Each section connects existing evidence to the empirical questions motivating the present study.

2.1. Hypothetical versus experiential attitudes

Surveys of AV attitudes consistently reveal widespread mistrust. Internationally, Kyriakidis et al. (2015) surveyed 5000 respondents across 109 countries and found that top concerns centered on software hacking, legal liability, and equipment failure. This is closely mirrored in the United States: Schoettle and Sivak (2014) reported that 26% of U.S. respondents were "very concerned" about equipment failure and unexpected performance, while consecutive national polls demonstrate that domestic distrust has remained persistent over several years (AAA Foundation, 2023, 2024, 2025). The point these surveys collectively establish is clear: prior to actual experience, the broader public shares a stable and multi-dimensional skepticism toward AV technology.

Yet much of this evidence is hypothetical. Three independent systematic reviews confirm that virtually no prior survey research had included respondents with actual AV experience (Becker & Axhausen, 2017; Gkartzonikas & Gkritza, 2019; Keszey, 2020). Highlighting this methodological gap, Tennant et al. (2019) reviewed 58 surveys (28 academic and 30 commercial) and explicitly questioned whether "attitudes to a hypothetical object tell us anything about how the public will respond when confronted by the technology itself". Even where controlled exposure has been attempted, it remains limited: Li et al. (2025)'s review of 99 pedestrian-AV interaction studies found that all relied on controlled methods (such as simulators or virtual reality), with a median sample size of 41 participants. The behavioral psychology literature provides strong theoretical grounds to doubt that such hypothetical attitudes will translate to actual behavior: stated intentions are unreliable predictors of real-world decisions, particularly for novel technologies where people lack direct experience against which to calibrate their expectations (Conner & Norman, 2022; Sheeran, 2002).

Field trials involving real vehicles have begun to emerge, but early studies were predominantly confined to controlled settings such as campus shuttles and low-speed segregated lanes (Mouratidis & Cobeña Serrano, 2021; Nordhoff et al., 2018; Paddeu et al., 2021; Salonen & Haavisto, 2019; Xu et al., 2018), which provide limited interaction with the mixed traffic that defines urban deployment. More recent robotaxi pilots in Guangzhou (Dai et al., 2021), Shanghai (Mao et al., 2025), and San Francisco (Power, 2024) operate in naturalistic conditions but remain observational. Isolating experience effects from pre-existing differences would require designs that measure trust immediately before and after first rides.

Even so, the evidence that does exist reveals a consistent disconnect between stated pre-exposure attitudes and actual post-exposure behavior. Power (2024) ($N = 3773$) documented a 56-percentage-point trust gap between actual robotaxi riders (76%) and the general population (20%). When behavioral choices are measured directly, the gap is even larger: Fink et al. (2025) found that 97% of participants (70 of 72, drawn from an initial survey of $N = 444$) willingly boarded an AV when one arrived, regardless of prior stated trust, including 33 of 34 who had explicitly stated they would not ride. Furthermore, the effects of experience appear to be selective rather than uniform. Mao et al. (2025), studying Shanghai robotaxi users, found that ride experience significantly reduced perceived risk and diminished the importance of driving enjoyment while leaving car ownership attitudes unchanged.

What remains unclear is how attitudes are structured among populations with sustained, naturalistic exposure across multiple interaction modes, particularly when that exposure takes different forms (riding, observing, sharing the road) that may provide feedback on different dimensions of trust.

2.2. Trust calibration and experience effects

As introduced previously, Lee and See's (2004) trust calibration framework explains how real-world interaction alters these perceptions. Because trust updates dynamically as people accumulate evidence, different types of AV exposure may engage different trust formation processes. Hoff and Bashir (2015) extended this into a three-layered model distinguishing *dispositional trust* (stable individual differences), *situational trust* (context-dependent variability), and *learned trust* (dynamic adjustments based on interaction).

This framework yields specific predictions. When observed performance meets or exceeds expectations, performance-based trust increases; when it violates expectations, trust erodes: often asymmetrically, as negative experiences tend to outweigh positive ones (Muir & Moray, 1996; Slovic, 1987). Process-based trust develops as users come to understand *how* a system works, not just *that* it works. Purpose-based trust reflects confidence that the system was designed with appropriate goals and is being deployed appropriately.

Applied to AVs, this framework suggests that different types of exposure may provide feedback on different trust bases. Passengers directly observe whether the vehicle performs reliably: whether it stops smoothly, drives safely, and arrives without incident. This experience provides rich performance feedback. Pedestrians and cyclists, by contrast, observe only external behavior: they see AVs yield or fail to yield, accelerate or brake, but have limited insight into internal reliability. Drivers sharing the road occupy an intermediate position, observing AV behavior in traffic but from outside the vehicle. Empirical evidence is consistent with this differential pattern (see Section 2.1; Mao et al. 2025).

Importantly, Lee and See's framework distinguishes *global trust* (overall confidence in the automation) from *calibrated trust* (trust that appropriately matches system capabilities across contexts). Well-calibrated trust is specific: a person might trust an AV to handle highway driving but not construction zones, or trust the sensors but worry about software updates. This distinction motivates our examination of specific concern types rather than global trust alone: we ask whether passengers report higher trust, as well as whether their concern patterns differ systematically from those of non-passengers.

Meanwhile, concerns about how *other humans* behave around AVs lie outside the performance feedback loop: no amount of riding experience informs expectations about human behavior in mixed traffic. This asymmetry generates a theoretical prediction: passenger experience should attenuate technology-related concerns through direct performance feedback, while human-behavior concerns should remain elevated and stable across all interaction modes regardless of exposure type.

The selection problem

Cross-sectional comparisons of riders and non-riders cannot establish whether experience causes trust or whether trusting individuals select into ridership. Early adopters of new technologies tend to be more optimistic, risk-tolerant, and technologically sophisticated (Rogers, 2003), raising the possibility that the rider-trust association reflects pre-existing dispositions rather than a treatment effect (learning from experience).

Most existing studies acknowledge this limitation. Fink et al. (2025) note that their controlled, university-based sample limits generalizability. Li et al. (2025) highlight representativeness concerns with convenience samples. The fundamental challenge is that we cannot observe the counterfactual: what would a rider's trust have been had they not ridden?

Several indirect signals can help triangulate. First, a *dose-response* pattern (rising trust with cumulative rides) would suggest experience effects operating beyond initial selection, since selection determines *whether* someone rides but not *how many times* they continue. Second, and more novel, is what we term the *ambient exposure* signal: in an environment saturated with AVs, the decision *not* to ride becomes informative. If even non-riders in AV-dense cities show high trust relative to national baselines, this suggests that ambient exposure (i.e. simply living among AVs) shapes attitudes independently of ridership decisions. Liu et al. (2024)'s revealed preference study in Lake Nona, Florida provides suggestive evidence: after four years of autonomous shuttle operations, trust in safety showed minimal effect on use intentions, while perceived usefulness and usability dominated. The authors interpret this as indicating that "as users become more familiar with AS, they prioritize essential mobility aspects…over technological novelty", consistent with trust becoming calibrated and receding as a decision factor.

The non-rider in San Francisco has had ample opportunity to ride but has not; their trust level reflects something other than unfamiliarity.

These considerations motivate our study design. By surveying residents of San Francisco, we can examine both dose-response patterns among passengers and baseline trust among non-riders, even without making strong causal claims.

2.3. The structure of safety concerns

Most AV perception research focuses on global attitudes: Does the public trust AVs? Are people willing to ride? Yet people may trust AVs in some respects while being concerned about others. Understanding what people worry about matters for both theory and policy.

Prior research has established that AV-related concerns span multiple dimensions. Beyond the system failure and hacking concerns documented in international surveys (Section 2.1), Keszey (2020)'s systematic review identified perceived risk as the most consistent negative predictor of behavioral intention, while Gkartzonikas and Gkritza (2019) noted that safety concerns appear across virtually all stated preference research. The AAA Foundation's national surveys found that technology malfunction consistently ranks among top worries, alongside driver over-reliance, loss of control, and cybersecurity (Kim & Kelley-Baker, 2021).

What remains underexplored is whether concern patterns vary by *mode of interaction*. Does riding in an AV recalibrate specific concerns while leaving others intact? Do pedestrians and cyclists, who experience AVs from outside, hold different concerns than passengers? Lee and See's framework suggests they might: passengers receive performance feedback about technology reliability, potentially reducing technology-related concerns, but they receive no feedback about how *other humans* will behave in mixed traffic. Pedestrians, who must predict both AV and human-driver behavior, may hold different concerns altogether.

This gap motivates our analysis of concern types across interaction modes.

2.4. Beyond passengers: Multi-modal road users

The vast majority of AV perception research adopts a passenger-centric perspective, asking whether people would ride in or trust AVs as occupants. Yet as Tennant et al. (2019) observe, "for most the first encounter will be sharing the road with rather than using an AV". This matters because external encounters differ fundamentally from passenger experiences.

Passengers observe AV capabilities from a position of delegated control: they have chosen to ride, they benefit directly when the ride goes well, and they experience the technology from inside the vehicle. Pedestrians and cyclists, by contrast, are vulnerable road users who must trust that the AV will detect and yield to them, without the implicit communication channels that typically govern human-driver interactions.

Communication is central to this concern. Rasouli and Tsotsos (2020) document that pedestrians routinely establish eye contact with drivers to confirm they have been seen, and that drivers who make eye contact are more likely to yield at crosswalks. AVs eliminate the driver from this exchange. While some evidence suggests pedestrians treat vehicles as entities rather than engaging with drivers (Dey et al., 2017), the absence of a visible operator introduces uncertainty about whether one has been detected and how the vehicle will behave. Some AV operators have attempted to address this gap through external signals. Waymo's vehicles, for instance, display an indicator showing when the vehicle has detected a pedestrian and intends to yield. Whether such signals effectively substitute for eye contact in complex urban traffic remains an open question.

Empirical research on non-passenger AV perceptions is limited and unevenly distributed. Li et al. (2025)'s systematic review of 99 pedestrian-AV interaction studies found that 92% focused on pedestrians (only 8% on cyclists), and nearly all used controlled methods rather than naturalistic observation. Field studies that do exist were typically conducted in low-speed campus or shuttle environments (Madigan et al., 2017; Mao et al., 2025; Rasouli & Tsotsos, 2020; Rothenbücher et al., 2016), with De Ceunynck et al. (2022) providing a rare exception through behavioral observation of pedestrians, cyclists, and e-scooter users on public roads in Oslo.

Survey research that examines perceptions across multiple road-user types within a single sample is a growing but still small literature. Hulse et al. (2018) provided the foundational treatment, surveying 916 UK residents who rated safety hypothetically as drivers, passengers, pedestrians, and cyclists. Penmetsa et al. (2019) extended the design to Pittsburgh during the early Uber AV-testing period, analyzing 798 general-public responses to a survey administered by the Bike PGH advocacy organization. Pyrialakou et al. (2020) surveyed 400 Phoenix-area residents about hypothetical safety while driving, cycling, and walking near AVs using three ordered probit models. Rahman et al. (2023) subsequently analyzed two waves of Pittsburgh BikePGH data ($N = 1{,}120$ in 2017, $N = 795$ in 2019) to track how perceptions shifted as AV visibility increased. Li et al. (2023) provides the most recent large multi-modal survey, recruiting 1,205 Queensland residents (456 pedestrians, 339 cyclists, 410 drivers) within a Theory of Planned Behavior framework to hypothetical scenarios about fully automated vehicles.

Modal coverage within this literature is uneven. Pedestrians and cyclists appear in every study, drivers in roughly half, and AV passengers only in Hulse et al. (2018), and only as a hypothetical role for participants without ridership experience. The literature also shares the sample-level features discussed in Section 2.1: online recruitment, AV exposure measured indirectly rather than as a screening criterion, and data collected before commercial robotaxi deployment. The closest peer to the present study is Nordhoff et al. (2025), who conducted semi-structured interviews with 52 U.S. residents screened for regular driverless AV interactions as pedestrians, cyclists, motorcyclists, or car drivers; its qualitative scale and the absence of AV passengers from its screener leave the quantitative multi-modal question open in a naturalistic deployment context.

The present study addresses this multi-modal coverage gap. We surveyed 811 Bay Area residents during May–June 2025, by which time Waymo's commercial robotaxi service in San Francisco had matured to thousands of paid rides a day (Commission, 2025). All respondents were required to have direct AV experience as a passenger, pedestrian, cyclist, or driver. To our knowledge, this is the first quantitative study to examine AV passengers alongside all three non-passenger road-user types within a single sample drawn from a naturalistic and large-scale commercial operation setting.

2.5. Human unpredictability in mixed traffic

Mixed-traffic environments introduce strategic uncertainty that shapes both behavior and perceptions. Millard-Ball (2018) formalizes pedestrian-vehicle encounters as "crosswalk chicken": behavior depends on beliefs about the other party's willingness to yield. Human drivers introduce uncertainty, since they might be distracted or aggressive, which makes pedestrians behave with caution. AVs, programmed to be risk-averse, could alter this risk calculation: knowing an AV will yield, pedestrians might exploit this and, in turn, behave less conservatively.

Yet this framing captures only half the uncertainty. In mixed traffic, *all* road users face unpredictability about human behavior: not just AVs facing pedestrians, but pedestrians facing human drivers, passengers uncertain whether others will exploit AV caution, and drivers unsure how pedestrians will respond to nearby AVs. This multi-directional uncertainty reinforces the prediction derived in Section 2.2: because human behavioral unpredictability lies outside any individual's performance feedback loop, concerns about it should persist across all interaction modes even as technology-related concerns attenuate.

Early field evidence on behavioral adaptation remains limited to low-speed campus settings (Rasouli & Tsotsos, 2020; Rothenbücher et al., 2016). Whether pedestrians in mature urban deployments exploit AV predictability or instead exercise greater caution remains unclear. Our data addresses both questions: how concern patterns vary by exposure type, and how pedestrians report behaving around AVs in naturalistic conditions.

2.6. Summary

The existing literature leaves three gaps central to understanding AV trust and safety perceptions in real deployment contexts:

1. **Hypothetical versus experiential data:** Most evidence comes from respondents who have never encountered an AV, and the few studies examining experienced respondents reveal substantial attitude discrepancies (Section 2.1).
2. **Limited multi-modal coverage:** Research mostly frames AV acceptance from a hypothetical-passenger perspective; the smaller body of multi-modal work concentrates on pedestrians and cyclists, with drivers as a secondary group and actual AV passengers almost never sampled alongside non-passengers (Section 2.4).
3. **Global trust versus concern type:** Studies typically measure overall trust rather than specific concern dimensions. Existing frameworks predict that experience might recalibrate technology-related concerns while leaving human-behavior concerns intact, but this has not been examined in naturalistic deployment.

3. Methodology

This section describes our data collection procedures and analytic approach. Section 3.1 details the survey design, sampling strategy, and measures. Section 3.2 presents statistical methods for addressing our research questions.

3.1. Data collection

We conducted intercept surveys of 811 Bay Area residents during May-June 2025. A professional data collection firm with extensive San Francisco experience recruited participants across 18 distinct neighborhood areas spanning the city. Sampling locations encompassed commercial and retail corridors, parks and public spaces, transit hubs, office districts, and residential neighborhoods. These locations spanned neighborhoods with American Community Survey (2019–2023) median household incomes ranging from approximately $65,000 to $229,000 (see Appendix D for survey locations, neighborhood-level detail, and methodology). While we did not collect individual-level socioeconomic data, the geographic distribution of sampling locations covers a wide range of neighborhood contexts across the city. Surveys were administered during morning and afternoon time blocks to capture commuters, shoppers, and residents engaged in routine activities.

Eligibility required: (1) age 18 or older, (2) Bay Area residence, and (3) at least one direct AV interaction: riding as a passenger, encountering AVs as a pedestrian or cyclist, or sharing the road as a driver. Interviewers trained on survey administration, informed consent, and neutral question delivery screened potential participants for eligibility and administered the survey via tablet. No identifying information was collected. No incentives for participation were provided.

Sampling strategy and classification We employed stratified convenience sampling with adaptive quotas to achieve broad representation by interaction mode. The survey first recorded *all* interaction modes for each respondent (passenger, pedestrian, cyclist, driver), then administered mode-specific questions based on an assigned primary classification, followed by a common set of questions administered to all respondents. Complete survey logic and branching structure appear in the survey instrument (see Data availability).

For analytic purposes, we assigned each respondent to a mutually exclusive primary interaction category. Given uncertainty about passenger prevalence and the theoretical importance of isolating riding experience, we prioritized passenger identification: respondents who had ever ridden in an AV were classified as *Passengers* (n = 376), regardless of whether they also reported pedestrian,

Table 1
Sample characteristics.

Characteristic	Full Sample n	%	Passengers (n = 376)	Ped/Cyclists (n = 314)	Drivers (n = 121)	p
Total	811	100.0				
Residence						
San Francisco	694	85.6	90.7%	84.7%	71.9%	<.001[a]
Elsewhere in Bay Area	117	14.4				
Gender						
Male	436	53.8	54.0%	53.5%	53.7%	0.99[a]
Female	375	46.2				
Age, mean (SD)	44.6 (16.9)		40.4 (15.7)	48.0 (17.4)	48.5 (16.1)	<.001[b]
Age group						
18–24	85	10.5				
25–34	190	23.4				
35–44	164	20.2				
45–54	108	13.3				
55–64	125	15.4				
65+	132	16.3				
AV encounter frequency						
Daily	503	62.0	67.6%	61.1%	47.1%	<.001[a]
2–3 times/week	179	22.1				
Weekly	81	10.0				
2–3 times/month	41	5.1				
Monthly or less	6	0.7				

Note: $N = 811$. Age missing for 7 respondents. Group comparison columns show % or mean (SD) within each interaction type. [a]Chi-square test. [b]One-way ANOVA. Passengers are significantly younger and more likely to be San Francisco residents than non-passengers. Gender does not significantly differ across groups.

cyclist, or driver exposure. Among non-passengers, we classified respondents as *Pedestrians/Cyclists* ($n = 314$) or *Drivers* ($n = 121$) based on their primary reported interaction mode.

This classification strategy was chosen to address two goals: (1) isolating riding experience as distinct from other exposure types, and (2) oversampling pedestrians and cyclists, who represent an understudied perspective in the AV literature despite their physical vulnerability (Das, 2021). In the initial phase of data collection, respondents reporting multiple non-passenger interaction modes were classified as pedestrians/cyclists. After achieving $n > 250$ pedestrians/cyclists, we implemented a soft quota to ensure at least 100 drivers were captured, given the importance of understanding how other vehicle operators perceive AVs in mixed traffic.

In a commercial urban deployment of this magnitude, multi-modal exposure is the norm: almost all AV passengers also reported encountering AVs as a pedestrian, cyclist, or driver, so a "rider-only" category was effectively absent from our sample and is unlikely to exist in any comparable deployment. Appendix B confirms that trust rates are substantively unchanged when multi-mode respondents are classified by self-reported primary mode, when overlapping categories are allowed, or when multi-mode respondents are excluded entirely.

Sample demographics broadly align with available San Francisco population parameters on the variables we measured. Gender composition (53.8% male) closely matches U.S. Census estimates (51.4%), and the proportion of respondents aged 65 and older (16.3%) is slightly lower than the citywide rate (17.2%; U. S. Census Bureau 2024b). Working-age adults (25–64) are somewhat over-represented relative to Census figures (73.1% in our sample versus 62.9% citywide), consistent with intercept sampling in public spaces during daytime hours. Table 1 summarizes the demographic composition of the full sample and each interaction subgroup. Passengers are significantly younger than non-passengers (mean age 40.4 vs. 48.0–48.5 years; ANOVA $F = 22.30$, $p < 0.001$) and more likely to be San Francisco residents (90.7% vs. 71.9–84.7%; $\chi^2 = 26.49$, $p < 0.001$). Gender composition does not differ significantly across groups ($\chi^2 = 0.02$, $p = 0.99$).

3.1.1. Variables collected

Trust. We asked: "Are you afraid of AVs, or do you trust them to be safe?" with response options: Trust / Unsure / Afraid. This wording parallels AAA Foundation national surveys (AAA Foundation, 2024), enabling illustrative comparison to nationally representative samples of predominantly unexposed respondents, though differences in sampling frames preclude strict equivalence. For statistical analysis, we coded responses as trusting (1) versus unsure or afraid (0). This binary coding is conservative: collapsing "unsure" with "afraid" treats ambivalence as distrust, but facilitates comparison with prior work and avoids over-interpreting distinctions that may not be meaningful to respondents.

Safety concerns. Respondents selected concerns from six options: "AVs struggling with unpredictable human behavior", "lack of AV safety performance information", "system malfunctions or software errors", "cybersecurity (hacking, privacy)", "AVs not being properly regulated", "AV response to emergency vehicles", plus an option to report "No concerns". Multiple selections were permitted.

Question wording appears in the survey questionnaire (see Data Aavailability). These categories were informed by concerns identified in prior large-scale surveys of AV perceptions (Kim & Kelley-Baker, 2021; Kyriakidis et al., 2015).

Close calls. Non-passengers indicated whether they had experienced “a close call or unexpected interaction with an AV” (binary). This subjective measure captures perceived rather than objective safety events, but perceptions are themselves consequential for attitude formation (Slovic, 1987).

Passenger experience. Passengers reported: number of rides taken (open-ended numeric), perceived safety rating (1 = very unsafe to 5 = very safe), and, for repeat riders, whether they currently feel “more safe,” “about the same,” or “less safe” compared to their first ride.

Covariates. Age, gender (male/female/prefer not to say), San Francisco residence (binary: San Francisco city limits vs. elsewhere in Bay Area), and AV exposure indicators by mode.

Additional measures. The survey included questions on behavioral adaptations (e.g., whether pedestrians act more cautiously around AVs), comparative safety ratings (AVs vs. human-driven vehicles), and desired safety improvements. The complete instrument appears in the repository (see Data availability).

This study was approved by UC Berkeley’s Institutional Review Board. All participants provided verbal informed consent.

3.2. Analysis

Our analysis is organized into four parts: (1) How does trust among AV-exposed residents compare to national baselines and vary by interaction mode? (2) What is the magnitude of the passenger-trust association, and how much might selection bias account for it? (3) Among passengers, does cumulative experience predict trust? (4) How do close calls and safety concerns shape perceptions?

Heteroskedasticity-robust standard errors (HC1) are reported throughout. We describe our analytic approach for each part below; supplementary technical details appear in Appendices A and C.

3.2.1. Trust differences by exposure and mode

A one-sample z-test was selected to compare our sample proportion to the national benchmark, as this test is appropriate when comparing an observed proportion against a known population value with a sufficiently large sample.

Chi-square tests of independence were used to assess trust differences across interaction modes, as they are appropriate for comparing proportions across mutually exclusive categorical groups without distributional assumptions.

Odds ratios (OR) are reported as effect size measures given the binary outcome variable.

Sensitivity to selection: The passenger-trust association does not establish causality: riding may increase trust, or individuals with higher baseline trust may select into ridership. We first assess whether the association is robust to observable confounders by estimating regression models (OLS and logit) controlling for age, gender, and AV exposure, and by propensity-score matching that pairs passengers with demographically similar non-passengers on these covariates plus San Francisco residence. To further probe selection on unobservables, we employ instrumental variables (IV) estimation using San Francisco residence as an instrument, exploiting differential access to robotaxi service due to supply-side geographical constraints. This strategy has significant limitations (the exclusion restriction is questionable and confidence intervals are wide), and therefore we interpret IV estimates as suggestive rather than definitive. Full specification and robustness checks appear in Appendix A.

3.2.2. Within-passenger experience effects

Among passengers, we examine whether cumulative experience predicts safety perceptions. The central challenge is distinguishing genuine learning (experience increases trust) from survivorship bias (dissatisfied riders stop riding, leaving a positively selected sample of continuing riders). Our cross-sectional design cannot definitively separate learning from selection, so we adopt a triangulation strategy that accumulates evidence from multiple angles.

Pre-survivorship benchmark: We first compare safety ratings and trust between first-time and repeat riders using independent-samples t-tests (Cohen’s d for effect size). First-time riders provide a critical benchmark: they have self-selected into AV ridership but have not yet been subject to selective continuation. If first-time riders already show elevated trust relative to non-passengers, this is an association with experience occurring before survivorship can operate.

Dose-response pattern Among repeat riders, we compute Spearman correlations between number of rides and safety ratings, and present trust rates by ride-count category. A monotonic increase is consistent with both learning and survivorship; subsequent analyses probe which interpretation better fits the data.

Retrospective change reports Repeat riders reported whether they felt “more safe,” “less safe,” or “about the same” compared to their first ride. This provides quasi-longitudinal evidence: if the dose-response pattern was purely selection (always-trusting individuals continuing to ride), we would expect “no change” to dominate. A substantial proportion reporting increased safety is consistent with genuine recalibration. We also examine whether those reporting decreased safety have systematically fewer rides, as survivorship would predict.

Sensitivity analysis As an additional robustness check, we estimate a Heckman two-step selection model (Heckman, 1979); full specification appears in Appendix C. However, we lack a valid exclusion restriction, so identification relies on distributional assumptions (bivariate normality) that we cannot verify. We therefore treat the Heckman results as supplementary to the triangulation evidence above rather than as the primary basis of our analysis.

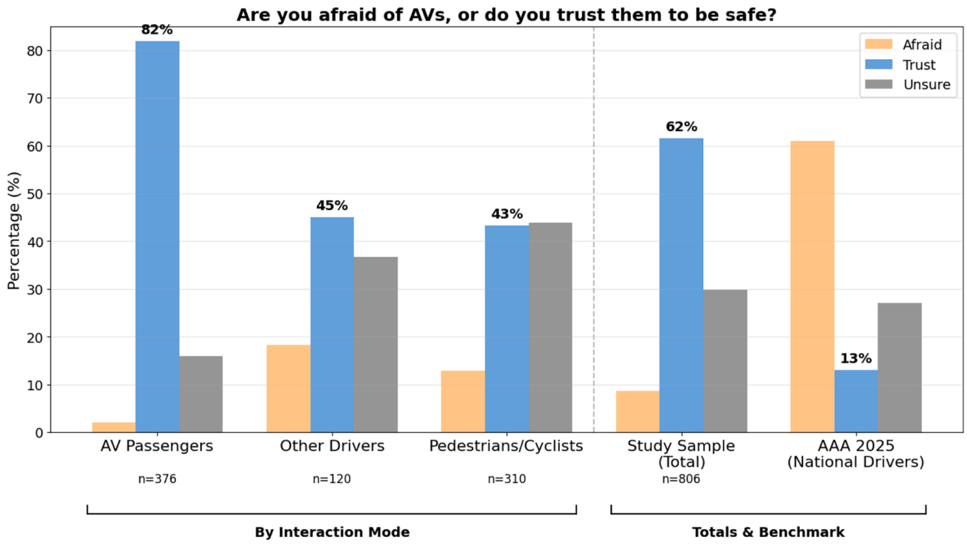

Fig. 1. Trust in AV safety rates by interaction mode. Subgroups reflect categories based on respondents' primary mode of AV interaction. Study Sample (Total) shows the pooled result across all respondents (N = 806). The AAA 2025 national driver survey (AAA Foundation, 2025), included for comparison, used identical question wording.

3.2.3. Close calls and safety concerns

For pedestrians/cyclists and drivers separately, we tested whether close call experience is associated with trust using chi-square tests, reporting odds ratios with 95% confidence intervals. Chi-square tests are appropriate here as both variables are binary and the test makes no distributional assumptions.

All respondents selected their AV safety concerns from seven options; multiple selections were permitted. For each concern, we calculated the percentage of respondents in each interaction mode who selected it and used chi-square tests to assess whether these percentages differed significantly across modes. To compare the two leading concerns, we used McNemar's test for paired proportions, which accounts for the non-independence of observations when the same individuals could select multiple concern categories.

Given multiple comparisons across concern types, we applied Benjamini-Hochberg false discovery rate (FDR) correction (Benjamini & Hochberg, 1995) to the family of concern-by-mode chi-square tests, with $\alpha = 0.05$. We report both uncorrected p-values and whether each test survives FDR correction.

4. Results

We present findings in four parts: trust differences by exposure mode, the passenger-trust association and selection bias, within-passenger experience effects, and close calls with safety concern patterns.

4.1. Trust differences by exposure and mode

Overall, 61.5% of respondents reported trusting AVs to be safe: 4.7 times higher than the 13% national rate reported by AAA Foundation (2025) ($z = 28.32$, $p < 0.0001$).

Trust varied significantly by interaction mode ($\chi^2(2) = 123.75$, $p < 0.0001$). Passengers reported 81.9% trust, while pedestrians/cyclists (43.2%) and drivers (45.0%) reported similar, substantially lower rates (Fig. 1). Pairwise comparisons confirmed that passengers differed significantly from both pedestrians/cyclists ($\chi^2 = 109.30$, $p < 0.0001$, OR = 5.95) and drivers ($\chi^2 = 61.01$, $p < 0.0001$, OR = 5.54). Pedestrians/cyclists and drivers did not differ from each other ($\chi^2 = 0.05$, $p = 0.82$). These unadjusted differences are robust to demographic adjustment: Section 4.2 presents estimates of the gap between passengers and non-passengers controlling for age, gender, and AV exposure via regression and propensity score matching, with functional-form robustness in Appendix A.

Even non-passengers in our sample reported markedly higher trust (43-45%) relative to the national baseline (13%). This difference suggests that ridership selection alone cannot explain this gap, but we cannot distinguish an ambient exposure effect from systematic differences between San Francisco residents and national samples in tech adoption, risk tolerance, and other unmeasured attributes (see Section 5).

4.2. Sensitivity of the passenger-trust association

Table 2 presents OLS and IV estimates of the passenger-trust association.

In naive OLS, passengers are 39 percentage points more likely to trust AVs than non-passengers in OLS ($\beta = 0.387$, $p < 0.001$). This estimate is likely biased upwards if individuals with higher baseline trust select into ridership. The IV point estimate is smaller

Table 2
Passenger-trust association: OLS, matching, and instrumental variables estimates.

Model	Estimate	Robust SE	95% CI	*p*-value
OLS	0.387	0.032	[0.325, 0.449]	<.001
Propensity score matching (ATT)	0.355	0.033	[0.291, 0.418]	<.001
2SLS (IV)	0.314	0.215	[-0.108, 0.736]	0.144
Instrument diagnostics				
First-stage F	28.7		(threshold > 10)	
Endogeneity test			$p = .71$	

Note: $N = 799$. Dependent variable is binary trust indicator. OLS controls for age, gender, and non-passenger AV exposure. IV instruments passenger status with San Francisco residence. PSM is the average treatment effect on the treated from 1-to-1 nearest-neighbor matching on the estimated propensity score (covariates: age, gender, San Francisco residence, AV exposure; caliper 0.05), with bootstrap standard errors (1,000 replications). Heteroskedasticity-robust (HC1) standard errors. See Appendix A for specification details and robustness checks.

Table 3
Trust and safety ratings by number of rides (passengers only).

Number of Rides	N	Mean Safety (1–5)	Trust Rate
1 (First-time)	91	4.14	70.3%
2	43	4.37	74.4%
3–5	92	4.47	80.4%
6–10	78	4.72	85.9%
11 +	72	4.76	98.6%

($\beta = 0.314$), but the confidence interval (-0.108, 0.736) is wide, spanning values consistent with no selection through substantial selection, hence we cannot draw conclusions about selection magnitude from these estimates. The first-stage F -statistic (28.7) exceeds conventional thresholds for instrument strength (Stock & Yogo, 2002). A falsification test shows that among non-passengers, San Francisco residents (43.8%) and non-San Francisco residents (43.2%) report statistically indistinguishable trust ($p = .92$), consistent with a valid instrument though not definitive. We present these results not as causal identification but as a specification check: if the passenger-trust association were entirely an artifact of selection or model dependence, we would expect it to attenuate substantially or reverse under IV assumptions. It does not.

An additional specification addresses confounding by observable demographics. Propensity score matching, which pairs each passenger with a demographically similar non-passenger on age, gender, San Francisco residence, and AV exposure, yields an average treatment effect on the treated of 0.355 (95% CI [0.291, 0.418]); covariate balance after matching is strong, with all standardized mean differences below 0.11 (down from 0.46 for age before matching; see Appendix A). That the matched estimate is slightly smaller than the unadjusted OLS estimate is consistent with passengers being younger and more likely to reside in San Francisco: adjusting for these differences shrinks the association but does not eliminate it.

The association is stable across functional forms (logit and probit), which is informative even without causal identification: meaning that the pattern we document is not fragile to standard econometric challenges (Altonji et al., 2005). Appendix A provides additional specification details and robustness checks.

4.3. Within-passenger experience effects

Among passengers, safety ratings and trust increased with cumulative rides. We applied the triangulation strategy described in Section 3.2.2 to assess whether this reflects learning, selection, or both.

Pre-survivorship benchmark. First-time riders reported significantly lower safety ratings than repeat riders: $M = 4.14$ (SD $= 0.77$) versus $M = 4.60$ (SD $= 0.69$); $t(374) = 5.32$, $p < 0.0001$, Cohen's $d = 0.64$. Trust rates showed the same pattern: 70.3% among first-time riders versus 85.6% among repeat riders. Notably, first-time riders already showed substantially elevated trust (70.3%) compared to non-passengers (43–45%), a 26 percentage-point difference occurring before any survivorship can operate.

Dose-response pattern. Among repeat riders, number of rides correlated positively with safety ratings (Spearman $\rho = 0.24$, $p < 0.0001$). Table 3 shows trust increasing monotonically from 74% (2 rides) to 99% (11 + rides).

The distribution of rides was right-skewed (median = 4, IQR = 2–10), reflecting a mix of casual users and frequent adopters.

Retrospective change reports. Among repeat riders, 37.9% reported feeling safer than on their first ride, 59.3% reported no change, and 2.8% reported feeling less safe. The small group reporting decreased safety ($n = 8$) had substantially fewer rides ($M = 4.2$) compared

Table 4
Trust rates by close call experience (non-passengers).

Group	Close Call		No Close Call		OR	p
	N	Trust	N	Trust		
Pedestrian/Cyclist	61	18.0%	249	49.4%	0.23	<.0001
Driver	30	26.7%	90	51.1%	0.35	0.034

Table 5
Safety concerns by interaction mode.

Concern	Passenger (N = 376)	Ped/Cyclist (N = 314)	Driver (N = 121)	χ^2	FDR
Unpredictable human behavior	52.4%	59.2%	54.5%	3.3	–
System malfunctions or software errors	45.2%	63.4%	53.7%	22.7***	✓
Emergency vehicle response	33.2%	45.9%	47.9%	14.7***	✓
Cybersecurity (hacking, privacy)	36.4%	40.1%	36.4%	1.1	–
Lack of regulation	22.3%	40.4%	42.1%	32.0***	✓
Lack of safety information	17.8%	35.0%	32.2%	28.2***	✓

Note: FDR column indicates whether test survives Benjamini-Hochberg correction at $\alpha = 0.05$. ***$p < 0.001$.

to those reporting increased safety ($M = 14.2$) or no change ($M = 10.7$), consistent with early attrition among those who develop negative perceptions.

Heckman selection model. The inverse Mills ratio was not significant ($\lambda = -0.54$, $p = 0.21$); the rides coefficient remained significant after correction ($\beta = 0.003$, $p = 0.018$). Full results appear in Appendix C.

Summary. Multiple lines of evidence suggest cumulative riding experience is associated with higher trust: (1) first-time riders show a substantial trust premium over non-passengers before survivorship operates; (2) over a third of repeat riders report feeling safer than initially; (3) dose-response estimates are stable to demographic controls; and (4) the Heckman correction does not alter findings. We cannot definitively separate learning from selection, but the pattern is consistent with experience-related changes operating alongside selective continuation.

4.4. Close calls and safety concerns

Close calls and trust. Among non-passengers, close calls were strongly associated with reduced trust (Table 4). Pedestrians and cyclists reporting a close call showed 18.0% trust versus 49.4% among those without such experiences ($\chi^2(1) = 18.38$, $p < 0.0001$, OR = 0.23 [0.11, 0.47]). Drivers showed a similar pattern: 26.7% trust with close calls versus 51.1% without ($\chi^2(1) = 4.49$, $p = 0.034$, OR = 0.35 [0.14, 0.90]). Overall, close calls are associated with 65–77% lower odds of trust.

Safety concern patterns. Table 5 presents safety concerns by interaction mode. The two most frequently reported concerns were “AVs struggling with unpredictable human behavior” (55.4%) and “system malfunctions or software errors” (53.5%), which did not differ significantly from each other overall (McNemar $\chi^2 = 0.70$, $p = 0.40$).

Concern patterns differed significantly by mode for four concerns after FDR correction (Table 5). Passengers were less likely to cite system malfunctions (45.2%) than pedestrians/cyclists (63.4%) or drivers (53.7%; $\chi^2(2) = 22.7$, $p < 0.0001$). Similar patterns emerged for lack of safety information ($\chi^2(2) = 28.2$, $p < 0.0001$), lack of regulation ($\chi^2(2) = 32.0$, $p < 0.0001$), and emergency vehicle response ($\chi^2(2) = 14.7$, $p = 0.0006$): in each case, passengers expressed less concern than non-passengers. In contrast, neither concern about unpredictable human behavior ($\chi^2(2) = 3.3$, $p = 0.19$) nor concern about cybersecurity ($\chi^2(2) = 1.1$, $p = 0.57$) varied by mode. The pattern is consistent with performance-based trust calibration: concerns that riding can directly inform (operational reliability, emergency response) attenuate with passenger experience, while concerns about unobservable risks (other humans' behavior, system security) persist regardless of exposure type. We emphasize that these patterns reflect observed associations within this specific deployment context and do not establish causal mechanisms: whether passenger experience directly causes concern attenuation or whether individuals with lower baseline technology concerns select into ridership cannot be determined from these cross-sectional data.

Behavioral adaptation. Among passengers, 72% rated AVs as safer than human-driven rides, with only 5% rating them less safe. Among pedestrians and cyclists, 46% act more cautiously around AVs than human-driven vehicles, while 17% trust AVs more and 37% treat them the same. Among drivers, 50% give AVs more space than human-driven vehicles, and 39% noted that AVs drive more predictably.

Over half of pedestrians and cyclists (54%) reported that inability to make eye contact with an AV makes them more hesitant or uncomfortable at crossings. Among those who behave more cautiously around AVs, 15% specifically cited lack of eye contact as a reason.

5. Discussion

This study documents trust and safety concern patterns among 811 San Francisco Bay Area residents with real-world AV exposure across a broad range of AV interaction modes. As a cross-sectional observational study, it documents associations rather than causal mechanisms. We interpret patterns through theoretical frameworks and propose hypotheses for experimental investigation.

5.1. The experience gap: ambient exposure and selection

As documented in Sections 1 and 2.1, national surveys consistently show low trust among unexposed populations, yet behavioral studies reveal substantial disconnects between stated attitudes and actual responses to AVs.

Mere exposure theory (Zajonc, 1968) predicts that repeated exposure to a stimulus increases positive affect, even without direct interaction. In AV contexts, this suggests that simply observing AVs navigate intersections, yield to pedestrians, and operate routinely could shift attitudes: an "ambient exposure effect" independent of ridership. Our finding that non-passengers in San Francisco report $3\times$ higher trust than unexposed national samples is consistent with this mechanism. Importantly, these non-passengers have had ample opportunity to ride in a mature market with extensive coverage but have chosen not to, sidestepping the selection critique that typically accompanies passenger-trust associations.

However, ambient exposure cannot be cleanly separated from residential selection. San Francisco residents may differ from nationally representative samples in unmeasured attributes that directly affect trust, so the $3\times$ trust gap could reflect ambient exposure effects, residential selection, or both operating simultaneously. Longitudinal studies tracking cities before and after AV deployment could isolate ambient exposure by measuring within-person trust changes as AVs arrive.

In the terms of the trust calibration framework (Section 2.2), these two possibilities attribute the elevated trust to different mechanisms. One explanation is a general urban exposure effect operating through *learned* trust: routine observation of AVs operating safely recalibrates trust through mere exposure (Zajonc, 1968) and observed performance feedback, a mechanism that should appear in any city with comparably mature deployment. A second is a San Francisco-specific effect operating through *dispositional* trust: the city's concentration of technology employment, its early adopter orientation (Rogers, 2003), and residents' potential social and professional ties to the firms developing AVs may raise baseline trust through channels unrelated to exposure itself. That national and cross-national surveys already document substantial geographic variation in AV attitudes (Kyriakidis et al., 2015; Schoettle & Sivak, 2014) is consistent with such a place-specific component. The two explanations are observationally equivalent in our cross-sectional data but carry different implications: under the first, elevated trust should emerge wherever AVs reach saturation; under the second, San Francisco is a best-case environment whose trust levels may not generalize. The density and volume of Waymo's San Francisco operation may further amplify a general exposure effect relative to nascent AV roll-outs. Telling the two apart requires comparative data across mature-deployment cities; existing single-city studies in other contexts (Dai et al., 2021; Liu et al., 2024; Mao et al., 2025) provide a foundation, but none yet measures non-rider trust across cities using a common instrument.

The passenger trust premium (82% vs. 43–45% non-passenger trust) is larger but equally ambiguous. Our sensitivity analysis using instrumental variables yields an estimate smaller than naive OLS, but the confidence interval is too wide to quantify selection. Crucially, the association remains positive and stable across specifications (see Appendix A). We cannot determine what proportion of the association reflects selection versus experience, but the pattern is not an artifact of model dependence.

Emerging evidence from other deployment contexts corroborates experience effects. Mao et al. (2025) found that ride experience in Shanghai increased AV adoption likelihood by 7.37%, with selective effects on specific attitudinal dimensions that parallel our concern profile results. Industry research similarly documents strong experience-trust links (Power, 2024).

These converging patterns from multiple deployment contexts remain observational and cannot establish causality. As discussed in Section 2.1, field trials range from controlled campus settings to naturalistic robotaxi pilots, but none yet incorporate the pre-post designs needed to isolate experience effects from pre-existing differences. Our cross-sectional findings provide the empirical grounding for such designs, and Section 5.4 proposes specific experimental strategies.

5.2. Concern recalibration through differential feedback

Lee and See's (2004) trust calibration framework outlined in Section 2.2 predicts that different exposure types should update different trust dimensions, but only when interaction provides relevant evidence.

Our findings reveal such differential belief updating. Passengers show markedly lower concern about system malfunctions (45%) than pedestrians/cyclists (63%), consistent with performance-based trust updating: riding provides direct evidence that AVs drive safely, stop smoothly, and complete trips successfully. This parallels findings from Technology Acceptance Model research (Davis, 1989; Venkatesh et al., 2003), where perceived usefulness updates through actual system use. This selective pattern is consistent with the Shanghai findings discussed above (Mao et al., 2025).

Notably, not all technology-related concerns varied by mode. Cybersecurity concerns (hacking, data privacy) remained constant across interaction types ($\chi^2 = 1.1,\ p = 0.57$). This distinction is theoretically coherent: riding provides direct evidence about operational performance (i.e. the vehicle responds to obstacles, completes trips, etc.) but no evidence about data security or system vulnerability to external attack.

Similarly, concern about unpredictable human behavior remains constant ($\sim 55\%$) across all interaction modes. This persistence makes theoretical sense: riding provides no information about how other road users will behave around AVs. Passengers observe pedestrians and drivers from inside the vehicle but cannot learn whether they will cross unpredictably or yield appropriately. The concerns that persist are those outside the performance feedback loop.

The multi-modal divergence is particularly revealing. Pedestrians and cyclists express high technology concerns despite regular AV exposure because they only observe behaviors externally (AVs yielding, accelerating, etc.) without insight into internal reliability. Their position denies them the performance feedback that would update technology concerns.

Meanwhile, all road users cite human behavior concerns, pointing to communication gaps in mixed traffic. Millard-Ball (2018) predicted that AVs' programmed caution creates exploitable yielding behavior. Yet our data show pedestrians acting more cautiously around AVs (46%), with over half (54%) citing eye contact absence as uncomfortable. This suggests communication breakdown rather than exploitation, at least in current deployment.

Li et al. (2025) conducted a review of 99 controlled experiments on pedestrian and cyclist interactions with AVs, documenting that eye contact and implicit communication cues are critical for establishing mutual understanding. The inability to "read" AV intentions through conventional signals creates uncertainty that increases caution rather than exploitation. Extensive literature (Bazilinskyy et al., 2021; Dey et al., 2020) documents attempts to restore pedestrian-AV communication through lights, displays, projections, and sounds, though effectiveness remains contested. Field studies comparing different designs show mixed results, suggesting that the optimal communication modality may be context-dependent (Li et al., 2025). Whether behavioral adaptation evolves toward Millard-Ball's exploitation prediction as familiarity increases remains an open question.

5.3. Trust asymmetry

Trust increased monotonically with cumulative rides (70% to 99%) but close calls were associated with sharp trust reductions (OR = 0.23-0.35). This asymmetry (slow building, rapid erosion) aligns with work on negativity bias in risk perception (Slovic, 1987) and prospect theory (Kahneman & Tversky, 1979), which document that losses loom larger than gains. For novel technologies, single alarming encounters may undo cumulative positive experiences.

However, directionality cannot be established. The pattern is consistent with close calls reducing trust (experience affecting attitudes) or with low baseline trust increasing close-call perception (attitudes affecting perception thresholds). The consistent pattern across pedestrians/cyclists and drivers suggests the association is meaningful, but experimental designs presenting equivalent AV behaviors to participants with measured baseline trust are required to establish causation.

If the erosion pathway were indeed causal and negative encounters disproportionately undermined trust, deployment strategies may need to prioritize minimizing alarming interactions over maximizing volume. A small number of close calls could offset widespread routine operation, particularly in early deployment when attitudes are still forming. However, if perception thresholds dominate, then transparency and communication about AV capabilities may matter as much as operational perfection.

5.4. Implications: hypotheses for future research

Our findings suggest testable hypotheses for future causal investigation:

H1: Ambient exposure effects: Non-passengers reported 3× higher trust than unexposed national samples, but residential selection cannot be ruled out. Longitudinal studies tracking cities pre- and post-AV arrival could test whether trust increases among non-adopters, isolating ambient exposure from selection.

H2: First-ride effects: First-time riders show elevated trust (70%) compared to non-passengers (43–45%), but this could reflect selection into first ride rather than the ride causing trust increase. Pre-post studies measuring trust immediately before and after individuals' first rides could isolate whether the experience itself shifts attitudes or whether those who choose to ride already differ (see Appendix A for discussion of identification strategies, including randomized ride credits and staggered rollout designs).

H3: Trust asymmetry: Close calls correlated with reduced trust, but directionality is unclear. Showing participants positive vs. negative AV encounters (smooth yielding vs. near-misses) could test whether single negative events disproportionately undermine trust built through multiple positive experiences.

H4: Communication modalities: Over half of pedestrians/cyclists in our sample cited inability to make eye contact as uncomfortable. While some AVs employ visual indicators, the effectiveness of different communication modalities remains unclear. Field experiments comparing designs (lights, displays, sounds, etc.) could identify which signals most effectively convey intent and reduce hesitation. Further, whether this discomfort attenuates with exposure or varies by demographic characteristics merits dedicated investigation.

These hypotheses require experimental designs that can establish causality where we document association. Converging evidence from Shanghai (Mao et al., 2025), Guangzhou (Dai et al., 2021), and industry studies (Power, 2024) suggests these patterns may generalize, though replication in cities with different cultural norms and traffic environments remains an important future research direction.

5.5. Conclusion

This study demonstrates that trust-experience relationships in autonomous vehicle contexts are mode-specific and not necessarily homogeneous. Different exposure types appear to provide feedback on different trust dimensions: direct performance experience (riding) is associated with lower technology concerns, while ambient exposure (observation) accompanies higher baseline trust, and human behavior concerns persist across all interaction modes regardless of experience type. These findings suggest that acceptance barriers may center less on perfecting technology than on managing communication gaps in mixed traffic.

Our multi-modal documentation of perception patterns within existing theoretical frameworks in a naturalistic saturation-level deployment provides empirical grounding for experimental designs that can establish causality where we document association. This study extends the empirical evidence base by showing how real-world exposure relates to specific, established dimensions of trust. A growing body of converging evidence suggests these patterns may generalize across deployment contexts. The path to broader acceptance may depend less on perfecting the technology than on understanding which trust dimensions shift with experience and which remain resistant to evidence, and on addressing the communication gaps that concern all road users regardless of exposure: a challenge that demands controlled investigation of the mechanisms our observational data cannot definitively establish.

5.6. Limitations

Several limitations of the present study merit emphasis. First, our cross-sectional design cannot establish causality: we cannot determine whether exposure increases trust or whether individuals with higher baseline trust select into exposure. The IV analysis provides a specification check but not causal identification. Second, the passenger-trust association remains partially ambiguous: while it is stable across specifications, we cannot quantify how much reflects selection versus experience. Third, directionality cannot be established for the close call-trust association; low baseline trust may increase the likelihood of perceiving interactions as close calls, rather than close calls reducing trust. Fourth, concern patterns are measured at a single point in time; longitudinal data would be needed to track how concerns evolve as deployment matures.

Fifth, and most critically, the intercept-survey format precluded the collection of individual-level socioeconomic data (income, education, race/ethnicity), which means we cannot disentangle ambient exposure effects from the demographic profile of San Francisco residents: a population that is relatively affluent, highly educated, and concentrated in technology industries. We cannot rule out that the trust gap between our sample and national baselines partly or wholly reflects this residential selection rather than exposure effects. Future research should collect income, education, and technology adoption measures (for example, through probability-based sampling designs), to allow explicit control of these confounders and enable population-level inference.

Sixth, the study was conducted during a period when Waymo operated as the sole commercial robotaxi provider in San Francisco. While our measures were framed around AVs generally, findings may partly reflect perceptions associated with Waymo's particular operational characteristics and may not generalize to other operators. For non-passengers, however, perceptions are formed through external encounters and may be less operator-specific. Replication across deployment contexts with multiple operators remains important.

Finally, our sample, while diverse, is not population-representative and findings should not be interpreted as population parameters. The intercept format may over-represent individuals who frequent public spaces and under-represent those with mobility constraints or privacy concerns.

CRediT authorship contribution statement

Carlos Guirado: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **SangHyouk Oum:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization; **Julia Griswold:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization; **Joan Walker:** Writing – review & editing, Methodology, Conceptualization.

Declaration of generative AI use

During the preparation of this work the author(s) used Claude Sonnet 4.5 in order to proofread and improve the grammar and clarity of some of the passages in this manuscript, and to assist with formatting. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no competing interests.

Acknowledgments

Funding for this study was provided by a grant from the California Office of Traffic Safety, through the National Highway Traffic Safety Administration. We would like to thank Matthew Raifman, Iman Mahdninia and Fiona Norton from the SafeTREC team for their feedback and support in survey design and data analysis.

Data availability

The complete dataset, analysis code, and survey instrument are available at https://github.com/tims-info/av-safety.

Supplementary material

Supplementary material associated with this article can be found in the online version at 10.1016/j.trf.2026.103738.

Appendix A. Robustness of the passenger-trust association to specification

This appendix provides technical details on the sensitivity analysis described in Section 3, including instrumental variables specification, identification assumptions and their limitations, and robustness checks.

A.1. Model specification

The first stage models passenger status as:

$$P_i = \gamma_0 + \gamma_1 Z_i + \gamma_2 A_i + \gamma_3 F_i + \gamma_4 E_i + \nu_i \tag{A.1}$$

The second stage substitutes predicted values:

$$T_i = \beta_0 + \beta_1 \hat{P}_i + \beta_2 A_i + \beta_3 F_i + \beta_4 E_i + \varepsilon_i \tag{A.2}$$

where T_i is a binary indicator for trusting AVs to be safe; P_i indicates passenger status (ever ridden as AV passenger); Z_i is San Francisco residence (the instrument); A_i is age in years; F_i is an indicator for female gender; and E_i indicates any non-passenger AV exposure (pedestrian, cyclist, or driver).

We assess instrument strength using the first-stage F-statistic ($F > 10$ indicates adequate strength; Stock and Yogo 2002). We test for endogeneity using a control function approach: including first-stage residuals in the outcome equation tests whether OLS is biased by selection on unobservables (Wooldridge, 2015).

A.2. Identification assumptions and limitations

This IV strategy requires three assumptions, each of which faces challenges in our setting:

Relevance. The instrument must predict passenger status. This assumption is satisfied: the first-stage F-statistic of 28.7 exceeds the Stock-Yogo threshold of 10, confirming that San Francisco residence strongly predicts ridership.

Exclusion restriction. The instrument must affect trust *only* through enabling ridership. This assumption is likely violated. San Francisco residents may differ from other Bay Area residents in tech adoption, baseline AV attitudes, exposure to AV-related media and discourse, or other unmeasured factors that could directly affect trust independent of ridership. While both populations are tech-forward relative to national samples, we cannot rule out attitudinal selection into San Francisco residence itself.

We probe the exclusion restriction through a falsification test. If the instrument is valid, San Francisco residence should not predict trust among non-passengers (for whom the ridership channel is inactive). Among non-passengers, San Francisco residents reported 43.8% trust versus 43.2% among non-San Francisco residents ($\chi^2 < 0.01$, $p = 0.92$). This null result is consistent with a valid instrument but cannot definitively rule out violations that operate only conditional on ridership.

Independence. The instrument must be independent of unmeasured confounders. This assumption is difficult to assess. If individuals who are predisposed to trust AVs differentially sort into San Francisco residence, the instrument would be correlated with the error term.

Statistical precision. A single binary instrument generates wide confidence intervals. The 95% CI for the IV estimate [-0.108, 0.736] includes zero, meaning we cannot statistically distinguish the IV estimate from either zero (selection explains everything) or the OLS estimate (selection explains nothing).

Endogeneity test power. The non-significance ($p = 0.71$) does not prove OLS is unbiased; it means we cannot detect bias with our sample size and instrument strength. Absence of evidence is not evidence of absence.

What would better identification require?. An ideal instrument would provide exogenous variation in ridership access uncorrelated with baseline attitudes. Candidates might include:

- *Randomized ride credits:* Lottery-based distribution of free ride vouchers to a representative sample, with trust measured before and after redemption.
- *Staggered geographic rollout:* Expansion of service boundaries creating sharp discontinuities in access among otherwise similar neighborhoods.
- *Waitlist experiments:* Random assignment to early versus delayed access among individuals who have expressed interest in trying AVs.
- *Workplace proximity:* Employment at firms located within versus outside service zones, among workers who did not choose their job based on AV access.

Table A.6
OLS coefficient stability.

	Restricted	Full
Passenger	0.379 (0.032)	0.387 (0.032)
Age	−0.000 (0.001)	0.000 (0.001)
Female	−0.108 (0.032)	−0.106 (0.032)
Any exposure	—	0.228 (0.150)
N	799	799
R^2	0.163	0.167

Note: Heteroskedasticity-robust standard errors in parentheses.

Table A.7
Functional form robustness: LPM, Probit, and Logit.

Functional form	Passenger estimate	SE
Linear probability model (OLS)	0.387	0.032
Probit (average marginal effect)	0.386	0.032
Logit (average marginal effect)	0.361	0.025

Note: $N = 799$. All specifications control for age, gender, and non-passenger AV exposure. Probit and logit report average marginal effects. The instrumental-variables estimate is similarly insensitive to functional form (IV-probit marginal effect = 0.311). Heteroskedasticity-robust (HC1) standard errors.

These instruments were unavailable in our cross-sectional design. Future research employing such designs could isolate experience effects from selection with greater precision than observational data permits.

Observable proxies for rideshare propensity (such as car ownership, existing rideshare app use) could serve as good controls, but likely fail the exclusion restriction required for instrumental variables, as they reflect demand-side behaviors correlated with AV attitudes.

A.3. Interpretation framework

Given these limitations, we interpret the IV results as follows:

- The *direction* of the IV-OLS divergence is informative: IV < OLS suggests positive selection (trusting individuals selecting into ridership), which aligns with theoretical expectations.
- The *magnitude* of divergence (approximately 20%) cannot be interpreted precisely; the wide confidence interval prevents conclusions about how much selection accounts for.
- The confidence interval including zero means we *cannot rule out* that selection explains the entire passenger-trust association.
- Given the questionable exclusion restriction, even a precise IV estimate would not establish causality.

We therefore treat the IV analysis as a specification check rather than causal identification: the passenger-trust association survives strict assumptions, suggesting it is not an artifact of model dependence. The triangulation evidence in Section 4.3 provides complementary evidence through a different analytic strategy.

A.4. Coefficient stability

Table A.6 presents coefficient stability in restricted and full OLS specifications.

Adding controls for non-passenger exposure changes the passenger coefficient by less than one percentage point, suggesting the estimate is not sensitive to observable confounders. However, coefficient stability with respect to observables does not guarantee stability with respect to unobservables (Altonji et al., 2005).

A.5. Functional form robustness

Table A.7 presents estimates from linear probability models (LPM), logit, and probit specifications. Probit and logit estimates are average marginal effects. Results are nearly identical across functional forms, indicating that findings are not sensitive to distributional assumptions.

Table A.8
Covariate balance before and after propensity-score matching.

Covariate	Pre-matching SMD	Post-matching SMD
Age	−0.460	0.005
Female	−0.009	0.022
San Francisco resident	0.283	−0.031
Any non-passenger exposure	−0.258	−0.105

Note: Standardized mean differences (passenger minus non-passenger), pooled standard deviation. $|\text{SMD}| < 0.10$ is a common threshold for adequate balance.

A.6. Propensity score matching

As a further check on confounding by observable demographics, we estimated the trust gap between AV passengers and non-passengers using propensity score matching. The propensity score (the estimated probability of being a passenger) was modeled via logistic regression on age, gender, San Francisco residence, and any non-passenger AV exposure. Each passenger was matched to the single non-passenger with the nearest propensity score (1-to-1 nearest-neighbor matching) within a caliper of 0.05.

Of 371 passengers with complete covariate data, 361 (97.3%) were matched within the caliper; common support spanned propensity scores from 0.149 to 0.677. Matching improved covariate balance. Before matching, passengers and non-passengers differed most on age (standardized mean difference, SMD $= -0.46$) and San Francisco residence (SMD $= 0.28$). After matching, all standardized mean differences fell below 0.11 and the age difference was effectively eliminated (Table A.8).

On the matched sample, passengers reported 82.3% trust versus 46.8% among matched non-passengers, an average treatment effect on the treated of 0.355 (bootstrap SE $= 0.033$ over 1,000 replications; 95% CI [0.291, 0.418]). This estimate is close to the OLS (0.387) and logit (0.361) estimates and within the range spanned by the IV point estimate, indicating that the passenger-trust association is not an artifact of demographic differences between passengers and non-passengers. As with the other specifications, matching adjusts only for observable covariates and cannot rule out selection on unobservables.

Appendix B. Robustness to classification approach

The main analysis assigns each respondent to a mutually exclusive interaction category, prioritizing passenger status (anyone who has ridden an AV is classified as "Passenger"). This appendix tests whether key findings are sensitive to alternative classification schemes.

We compare three approaches:

Assigned role (main specification): Priority classification as described in Section 3. Passengers receive priority; multi-mode non-passengers classified as pedestrian/cyclist or driver based on primary mode.

Pooled exposure: Allow overlapping categories. Respondents who report both pedestrian/cyclist and driver exposure are coded in both categories rather than forced into one.

Mutually exclusive: Exclude respondents with multiple non-passenger modes entirely. Compare only those with single-mode exposure.

Table B.9 presents trust rates under each approach. The key finding (that passengers report substantially higher trust than non-passengers regardless of specific non-passenger interaction mode) is robust across all specifications.

Table B.9
Trust rates by classification approach.

Classification	Pedestrian/Cyclist	Driver	χ^2	p
Assigned role	43.2% (N = 314)	45.0% (N = 121)	0.05	0.82
Pooled exposure	43.3% (N = 365)	44.5% (N = 229)	0.04	0.84
Mutually exclusive	42.9% (N = 206)	45.7% (N = 70)	0.18	0.91

Note: "Both" group (N = 159) shows 43.9% trust, consistent with single-exposure groups.

Trust rates are virtually identical across all specifications (42.9%–45.7%), and no comparison reaches statistical significance.

Appendix C. Heckman selection model: specification and sensitivity

As a sensitivity analysis to complement the triangulation approach described in Section 4.3, we estimated a Heckman two-step selection model (Heckman, 1979). This appendix documents the model specification, identification assumptions, and results (see Table C.10).

Table C.10
Heckman selection model results.

	Estimate	p-value
Selection equation (DV: Repeat rider)		
Intercept	- 0.42	0.18
Age	0.01	0.07
Female	- 0.15	0.31
Frequency	0.28	0.002
Outcome equation (DV: Safety rating)		
Intercept	4.12	<.001
Rides	0.003	0.018
Age	0.002	0.42
Female	- 0.08	0.29
Inverse Mills ratio (λ)	- 0.54	0.21

Note: $N = 376$ passengers; 285 repeat riders in outcome equation. Safety rated 1–5.

C.1. Model structure

The selection equation models the probability of becoming a repeat rider among all passengers:

$$\Pr(\text{Repeat}_i = 1) = \Phi(\gamma_0 + \gamma_1 \text{Age}_i + \gamma_2 \text{Female}_i + \gamma_3 \text{Frequency}_i) \tag{C.1}$$

where $\Phi(\cdot)$ is the standard normal CDF and Frequency is self-reported AV interaction frequency.

The outcome equation models safety ratings among repeat riders:

$$\text{Safety}_i = \beta_0 + \beta_1 \text{Rides}_i + \beta_2 \text{Age}_i + \beta_3 \text{Female}_i + \theta \lambda_i + \varepsilon_i \tag{C.2}$$

where λ_i is the inverse Mills ratio (IMR), computed as:

$$\lambda_i = \frac{\phi(\mathbf{Z}_i \hat{\gamma})}{\Phi(\mathbf{Z}_i \hat{\gamma})} \tag{C.3}$$

and $\phi(\cdot)$ is the standard normal PDF.

C.2. Interpretation

The coefficient θ on the IMR tests for selection bias. If $\theta \neq 0$, unobserved factors affecting the decision to continue riding also affect safety ratings, biasing naive estimates. A non-significant θ suggests selection may not be driving the observed dose-response pattern, though the test may be underpowered.

The coefficient β_1 on Rides, estimated conditional on the IMR, represents the association between cumulative experience and safety ratings after accounting for selection into repeat ridership.

C.3. Identification

The Heckman model is typically identified through an exclusion restriction: a variable that affects selection but not the outcome. We lack a theoretically justified exclusion restriction. Identification therefore relies on the assumption that errors in the selection and outcome equations follow a bivariate normal distribution, with nonlinearity in the IMR providing functional form identification.

This is a significant limitation. If the bivariate normality assumption fails (for example, if unobserved risk tolerance affects both continuation and ratings through a non-normal process), the correction may be unreliable. The test may also be underpowered given sample size constraints. We therefore present these results as one piece of supplementary evidence rather than as definitive identification of selection effects.

C.4. Results

The non-significant IMR ($\theta = -0.54$, $p = 0.21$) provides no statistical evidence that selection into repeat ridership biases the dose-response estimate. The rides coefficient remains significant after correction ($\beta_1 = 0.003$, $p = 0.018$).

Given the identification limitations noted above, we do not interpret the non-significant IMR as definitive evidence that survivorship bias is absent. Rather, we interpret these results as consistent with, but not proof of, the conclusion suggested by the triangulation evidence in Section 4.3: that experience-related changes likely operate alongside any selection, with selection alone unlikely to fully account for the observed dose-response pattern.

Table D.11
Interview location areas and neighborhood characteristics.

Neighborhood Area	Setting Type	Median HHI ($)
Civic Center/Hayes Valley	Transit/Office	65,000
North Beach	Commercial	83,000
Financial District[a]	Transit/Office	91,000
Russian Hill/Polk	Residential	112,000
Portola	Residential	119,000
Alemany Farmers Market	Market	131,000
Ingleside	Residential	131,000
Excelsior	Residential	131,000
Stonestown/West Portal	Commercial	136,000
Mission/Valencia	Commercial	152,000
Bernal Heights/Precita	Park/Residential	152,000
Golden Gate Park	Park	153,000
SoMa/Mission Bay	Transit/Office	154,000
Richmond/Clement St	Commercial	160,000
UCSF/Inner Sunset	Institutional	160,000
Dolores Park/Castro	Park/Commercial	197,000
Noe Valley/24th St	Commercial	198,000
Marina/Presidio	Park	229,000

Note: Median household income (MHI) from American Community Survey 5-Year Estimates, 2019–2023 (Table B19013), at the Zip Code Tabulation Area (ZCTA) level. When a neighborhood spans multiple ZCTAs, the reported value is the average of constituent medians. San Francisco citywide median household income is approximately $126,000 (ACS 2023). Interviews were conducted at multiple specific sites within each neighborhood area. [a]Financial District MHI reflects a small residential population (ZCTA 94104, MHI $49,896) averaged with the adjacent Embarcadero area (ZCTA 94111, MHI $131,343); the low residential count in 94,104 may not reflect the character of the daytime population interviewed at this location.

Appendix D. Sampling location characteristics

To characterize the socioeconomic diversity of our sampling locations, we mapped each interview site to a San Francisco neighborhood area and linked it to census-level income data. Median household income (MHI) figures are drawn from the U.S. Census Bureau American Community Survey (ACS) 5-Year Estimates (2019–2023), Table B19013, at the Zip Code Tabulation Area (ZCTA) level (U. S. Census Bureau, 2024a).[1]

Table D.11 presents the 18 neighborhood areas where interviews were conducted, classified by setting type and neighborhood-level MHI. Sampling locations ranged from lower-income areas such as Civic Center ($65,000) and North Beach ($83,000) to affluent neighborhoods including Noe Valley ($198,000) and Marina/Presidio ($229,000). The range reflects the economic heterogeneity of San Francisco, though we note that even the city's lower-income neighborhoods have median incomes above national averages, consistent with San Francisco's high overall cost of living.

These figures characterize the *neighborhoods* where interviews took place, not the socioeconomic status of individual respondents. Attributing neighborhood-level income to individuals interviewed in those locations is not possible, as respondents may not reside in or be representative of the areas where they were intercepted.

References

AAA Foundation (2023). AAA: Fear of self-driving cars on the rise. https://newsroom.aaa.com/2023/03/aaa-fear-of-self-driving-cars-on-the-rise/.

AAA Foundation (2024). AAA: Fear of self-driving cars persists as industry faces an uncertain future. https://newsroom.aaa.com/2024/03/aaa-fear-of-self-driving-cars-persists-as-industry-faces-an-uncertain-future/.

AAA Foundation (2025). AAA: Fear in self-driving vehicles persists. https://newsroom.aaa.com/2025/02/aaa-fear-in-self-driving-vehicles-persists/.

Altonji, J. G., Elder, T. E., & Taber, C. R. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of Catholic Schools. *Journal of Political Economy*, *113*(1), 151–184. https://doi.org/10.1086/426036

Bazilinskyy, P., Kooijman, L., Dodou, D., & de Winter, J. C. F. (2021). How should external human-machine interfaces behave? Examining the effects of colour, position, message, activation distance, vehicle yielding, and visual distraction among 1,434 participants. *Applied Ergonomics*, *95*, 103450. https://doi.org/10.1016/j.apergo.2021.103450

Becker, F., & Axhausen, K. W. (2017). Literature review on surveys investigating the acceptance of automated vehicles. *Transportation*, *44*(6), 1293–1306. https://doi.org/10.1007/s11116-017-9808-9

[1] Variable B19013_001E (Median Household Income in the Past 12 Months, in 2023 Inflation-Adjusted Dollars). When a neighborhood area spans multiple ZCTAs, the reported value is the simple average of the constituent ZCTA medians.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Commission, C. P. U. (2025). AV passenger service deployment - Quarterly reports. Quarterly submissions pursuant to D.20-11-046 (as modified by D.21-05-017) and D.24-11-002 https://www.cpuc.ca.gov/regulatory-services/licensing/transportation-licensing-and-analysis-branch/autonomous-vehicle-programs/quarterly-reporting.

Conner, M., & Norman, P. (2022). Understanding the intention-behavior gap: The role of intention strength. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.923464

Dai, J., Li, R., Liu, Z., & Lin, S. (2021). Impacts of the introduction of autonomous taxi on travel behaviors of the experienced user: Evidence from a one-year paid taxi service in Guangzhou, China. *Transportation Research Part C: Emerging Technologies*, *130*, 103311. https://doi.org/10.1016/j.trc.2021.103311

Das, S. (2021). Autonomous vehicle safety: Understanding perceptions of pedestrians and bicyclists. *Transportation Research Part F: Traffic Psychology and Behaviour*, *81*, 41–54. https://doi.org/10.1016/j.trf.2021.04.018

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

De Ceunynck, T., Pelssers, B., Bjørnskau, T., Aasvik, O., Fyhri, A., Laureshyn, A., Johnsson, C., Hagenzieker, M., & Martensen, H. (2022). Interact or counteract? Behavioural observation of interactions between vulnerable road users and autonomous shuttles in Oslo, Norway. *Traffic Safety Research*, *2*, 000008. https://doi.org/10.55329/fbhr3456

Dey, D., Habibovic, A., Pfleging, B., Martens, M., & Terken, J. (2020). Color and animation preferences for a light band eHMI in interactions between automated vehicles and pedestrians. In *Proceedings of the 2020 CHI conference on human factors in computing systems* CHI '20 (pp. 1-13). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3313831.3376325

Dey, D., Martens, M., Eggen, B., & Terken, J. (2017). The impact of vehicle appearance and vehicle behavior on pedestrian interaction with autonomous vehicles. In *Proceedings of the 9th international conference on automotive user interfaces and interactive vehicular applications adjunct* (pp. 158–162). https://doi.org/10.1145/3131726.3131750

Examiner, S. F. (2025). Waymo says its SF fleet has tripled in size since last public disclosure. https://www.sfexaminer.com/news/transit/waymo-finally-admits-there-are-way-more-robotaxis-in-sf-now/article_9c1d1d93-276a-4644-858a-3c29bc25b407.html.

Fagnant, D. J., & Kockelman, K. (2015). Preparing a nation for autonomous vehicles: Opportunities, barriers and policy recommendations. *Transportation Research Part A: Policy and Practice*, *77*, 167–181. https://doi.org/10.1016/j.tra.2015.04.003

Fink, P. D. S., Brown, J. R., Kutzer, K. M., & Giudice, N. A. (2025). Does trust even matter? Behavioral evidence for the disconnect between people's subjective trust and decisions to use autonomous vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour*, *114*, 99–117. https://doi.org/10.1016/j.trf.2025.05.024

Gkartzonikas, C., & Gkritza, K. (2019). What have we learned? A review of stated preference and choice studies on autonomous vehicles. *Transportation Research Part C: Emerging Technologies*, *98*, 323–337. https://doi.org/10.1016/j.trc.2018.12.003

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*(1), 153–161. https://doi.org/10.2307/1912352

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Hulse, L. M., Xie, H., & Galea, E. R. (2018). Perceptions of autonomous vehicles: Relationships with road users, risk, gender and age. *Safety Science*, *102*, 1–13. https://doi.org/10.1016/j.ssci.2017.10.001

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291. https://doi.org/10.2307/1914185

Keszey, T. (2020). Behavioural intention to use autonomous vehicles: Systematic review and empirical extension. *Transportation Research Part C: Emerging Technologies*, *119*, 102732. https://doi.org/10.1016/j.trc.2020.102732

Kim, W., & Kelley-Baker, T. (2021). Users' trust in and concerns about automated driving systems. Technical Report AAA Foundation for Traffic Safety Washington, D.C. https://aaafoundation.org/users-trust-in-and-concerns-about-automated-driving-systems/.

Kyriakidis, M., Happee, R., & de Winter, J. C. F. (2015). Public opinion on automated driving: Results of an international questionnaire among 5000 respondents. *Transportation Research Part F: Traffic Psychology and Behaviour*, *32*, 127–140. https://doi.org/10.1016/j.trf.2015.04.014

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Li, D., Mao, W., Pereira, F. C., Xiao, Y., Su, X., & Krueger, R. (2025). Analyzing the behaviors of pedestrians and cyclists in interactions with autonomous systems using controlled experiments: A literature review. *Transportation Research Part F: Traffic Psychology and Behaviour*, *114*, 270–307. https://doi.org/10.1016/j.trf.2025.05.031

Li, X., Kaye, S.-A., Afghari, A. P., & Oviedo-Trespalacios, O. (2023). Sharing roads with automated vehicles: A questionnaire investigation from drivers', cyclists' and pedestrians' perspectives. *Accident Analysis and Prevention*, *188*, 107093. https://doi.org/10.1016/j.aap.2023.107093

Liu, Y., Lu, K., Peng, Z.-R., & Zhai, W. (2024). Autonomous shuttle acceptance in an American suburban context: A revealed preference study in Lake Nona, Florida. *Travel Behaviour and Society*, *37*, 100865. https://doi.org/10.1016/j.tbs.2024.100865

Madigan, R., Louw, T., Wilbrink, M., Schieben, A., & Merat, N. (2017). What influences the decision to use automated public transport? Using UTAUT to understand public acceptance of automated road transport systems. *Transportation Research Part F: Traffic Psychology and Behaviour*, *50*, 55–64. https://doi.org/10.1016/j.trf.2017.07.007

Mao, H., Li, H., Jin, Z., Tu, H., & Hensher, D. A. (2025). The impact of ride experience on car travelers' adoption of private and shared autonomous vehicles. *Travel Behaviour and Society*, *41*, 101099. https://doi.org/10.1016/j.tbs.2025.101099

Millard-Ball, A. (2018). Pedestrians, autonomous vehicles, and cities. *Journal of Planning Education and Research*, *38*(1), 6–12. https://doi.org/10.1177/0739456X16675674

Mouratidis, K., & Cobeña Serrano, V. (2021). Autonomous buses: Intentions to use, passenger experiences, and suggestions for improvement. *Transportation Research Part F: Traffic Psychology and Behaviour*, *76*, 321–335. https://doi.org/10.1016/j.trf.2020.12.007

Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, *39*(3), 429–460. https://doi.org/10.1080/00140139608964474

NHTSA (2008). National motor vehicle crash causation survey: Report to Congress. Report DOT HS 811 059. U.S. Department of Transportation. crashstats.nhtsa.dot.gov.

Nordhoff, S., de Winter, J., Madigan, R., Merat, N., van Arem, B., & Happee, R. (2018). User acceptance of automated shuttles in berlin-schöneberg: A questionnaire study. *Transportation Research Part F: Traffic Psychology and Behaviour*, *58*, 843–854. https://doi.org/10.1016/j.trf.2018.06.024

Nordhoff, S., Hagenzieker, M., Lee, Y. M., Wilbrink, M., Merat, N., & Oehl, M. (2025). "It's just another car driving": Perceptions of U.S. residents interacting with driverless automated vehicles on public roads. *Transportation Research Part F: Traffic Psychology and Behaviour*, *111*, 188–210. https://doi.org/10.1016/j.trf.2025.01.024

Paddeu, D., Tsouros, I., Parkhurst, G., Polydoropoulou, A., & Shergold, I. (2021). A study of users' preferences after a brief exposure in a shared autonomous vehicle (SAV). *Transportation Research Procedia*, *52*, 533–540. https://doi.org/10.1016/j.trpro.2021.01.063

Penmetsa, P., Adanu, E. K., Wood, D., Wang, T., & Jones, S. L. (2019). Perceptions and expectations of autonomous vehicles—A snapshot of vulnerable road user opinion. *Technological Forecasting and Social Change*, *143*, 9–13. https://doi.org/10.1016/j.techfore.2019.02.010

Power, J. D. (2024). 2024 U.S. robotaxi experience study. https://www.jdpower.com/business/us-robotaxi-experience-study.

Pyrialakou, V. D., Gkartzonikas, C., Gatlin, J. D., & Gkritza, K. (2020). Perceptions of safety on a shared road: Driving, cycling, or walking near an autonomous vehicle. *Journal of Safety Research*, *72*, 249–258. https://doi.org/10.1016/j.jsr.2019.12.017

Rahman, M. T., Dey, K., Pyrialakou, V. D., & Das, S. (2023). Factors influencing safety perceptions of sharing roadways with autonomous vehicles among vulnerable roadway users. *Journal of Safety Research*, *85*, 266–277. https://doi.org/10.1016/j.jsr.2023.02.010

Rasouli, A., & Tsotsos, J. K. (2020). Autonomous vehicles that interact with pedestrians: A survey of theory and practice. *IEEE Transactions on Intelligent Transportation Systems*, *21*(3), 900–918. https://doi.org/10.1109/TITS.2019.2901817

Rogers, E. M. (2003). Diffusion of innovations. (5th ed.). New York: Free Press.

Rothenbücher, D., Li, J., Sirkin, D., Mok, B., & Ju, W. (2016). Ghost driver: A field study investigating the interaction between pedestrians and driverless vehicles. In *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)* (pp. 795–802). https://doi.org/10.1109/ROMAN.2016.7745210

Salonen, A. O., & Haavisto, N. (2019). Towards autonomous transportation. passengers' experiences, perceptions and feelings in a driverless shuttle bus in finland. *Sustainability*, *11*(3). https://doi.org/10.3390/su11030588

Schoettle, B., & Sivak, M. (2014). A survey of public opinion about autonomous and self-driving vehicles in the U.S., the U.K., and Australia, . https://cdn.theatlantic.com/assets/media/img/posts/UMTRI-2014-21.pdf.

Sheeran, P. (2002). Intention–behavior relations: A conceptual and empirical review. *European Review of Social Psychology*, *12*(1), 1–36. https://doi.org/10.1080/14792772143000003

Slovic, P. (1987). Perception of risk. *Science*, *236*(4799), 280–285. https://doi.org/10.1126/science.3563507

Stock, J. H., & Yogo, M. (2002). Testing for weak instruments in linear IV regression. Technical Working Paper 284. National Bureau of Economic Research. https://doi.org/10.3386/t0284

Tennant, C., Stares, S., & Howard, S. (2019). Public discomfort at the prospect of autonomous vehicles: Building on previous surveys to measure attitudes in 11 countries. *Transportation Research Part F: Traffic Psychology and Behaviour*, *64*, 98–118. https://doi.org/10.1016/j.trf.2019.04.017

U. S. Census Bureau (2024a). American community survey 5-year estimates, 2019–2023: Table b19013 (median household income in the past 12 months) by zip code tabulation area, San Francisco County, California. Variable B19013_001E. Retrieved via Census Bureau API. Accessed June 2025 https://api.census.gov/data/2023/acs/acs5.

U. S. Census Bureau (2024b). American community survey 5-year estimates, 2019–2023: Table DP05 (demographic and housing estimates) for San Francisco City, California. Accessed June 2025 https://data.census.gov/table/ACSDP5Y2023.DP05?g=160XX00US0667000.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425–478. https://doi.org/10.2307/30036540

Wooldridge, J. M. (2015). Control function methods in applied econometrics. *Journal of Human Resources*, *50*(2), 420–445. https://doi.org/10.3368/jhr.50.2.420

Xu, Z., Zhang, K., Min, H., Wang, Z., Zhao, X., & Liu, P. (2018). What drives people to accept automated vehicles? Findings from a field experiment. *Transportation Research Part C: Emerging Technologies*, *95*, 320–334. https://doi.org/10.1016/j.trc.2018.07.024

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, *9*(2, Pt. 2), 1–27. https://doi.org/10.1037/h0025848