eLife

Reviewed Preprint
v1 • June 24, 2024
Not revised

Neuroscience

Uncertainty-based causal inference modulates audiovisual temporal recalibration

Luhe Li ✉, **Fangfang Hong, Stephanie Badde, Michael S. Landy**

Department of Psychology, New York University • Department of Psychology, University of Pennsylvania • Department of Psychology, Tufts University • Center for Neural Science, New York University

Abstract

Cross-modal temporal recalibration is crucial for maintaining coherent perception in a multimodal environment. The classic view suggests that cross-modal temporal recalibration aligns the perceived timing of sensory signals from different modalities, such as sound and light, to compensate for physical and neural latency differences. However, this view cannot fully explain the nonlinearity and asymmetry observed in audiovisual recalibration effects: the amount of re-calibration plateaus with increasing audiovisual asynchrony and varies depending on the leading modality of the asynchrony during exposure. To address these discrepancies, our study examines the mechanism of audiovisual temporal recalibration through the lens of causal inference, considering the brain's capacity to determine whether multimodal signals come from a common source and should be integrated, or else kept separate. In a three-phase recalibration paradigm, we manipulated the adapter stimulus-onset asynchrony in the exposure phase across nine sessions, introducing asynchronies up to 0.7 s of either auditory or visual lead. Before and after the exposure phase in each session, we measured participants' perception of audiovisual relative timing using a temporal-order-judgment task. We compared models that assumed observers re-calibrate to approach either the physical synchrony or the causal-inference-based percept, with uncertainties specific to each modality or comparable across them. Modeling results revealed that a causal-inference model incorporating modality-specific uncertainty captures both the nonlinearity and asymmetry of audiovisual temporal recalibration. Our results indicate that human observers employ causal-inference-based percepts to recalibrate cross-modal temporal perception.

eLife assessment

In this **valuable** study, Li et al., set out to understand the mechanisms of audiovisual temporal recalibration - the brain's ability to adjust to the latency differences that emerge due to different (distance-dependent) transduction latencies of auditory and visual signals - through psychophysical measurements and modelling. The analysis supports a role for causal inference in recalibration, though the evidence is **incomplete**.

https://doi.org/10.7554/eLife.97765.1.sa3

1 Introduction

Perception is flexible and should change in response to perceptual error (Bedford, 1999). In a multimodal environment, systematic discrepancies between modalities indicate perceptual errors and thus the need for recalibration. Signals in different modalities from the same event can arrive with different physical and neural delays in the relevant brain areas (Fain, 2019; Pöppel, 1988; Spence & Squire, 2003). Cross-modal temporal recalibration has been considered a compensatory mechanism that attempts to realign the perceived timing between modalities to maintain perceptual synchrony across changes in the perceptual systems and the environment (reviewed by King, 2005; Vroomen and Keetels, 2010). This phenomenon is exemplified by audiovisual temporal recalibration, where consistent exposure to audiovisual asynchrony shifts the audiovisual temporal bias (i.e., point of subjective simultaneity) between auditory and visual stimuli in the direction of the asynchrony to which one has been exposed (Di Luca et al., 2009; Fujisaki et al., 2004; Hanson et al., 2008; Harrar & Harris, 2008; Heron et al., 2007; Keetels & Vroomen, 2007; Navarra et al., 2005; Roach et al., 2011; Tanaka et al., 2011; Vatakis et al., 2007, 2008; Vroomen & Keetels, 2010; Vroomen et al., 2004; Yamamoto et al., 2012).

The classic compensatory view is that recalibration serves to offset the physical and neural latency differences between modalities (Fujisaki et al., 2004), aiming for external accuracy, the agreement between perception and the environment (Zaidel et al., 2011). When external feedback is not available, recalibration may target internal consistency (Burge et al., 2010). In the context of the recalibration of relative timing, both theories predict similar behavior regardless of whether the goal is accuracy or consistency: the perceptual system will attempt to recalibrate for any amount of asynchrony so as to homeostatically restore a physical or perceived synchrony. Here, we formalize this hypothesis as the fixed-update model.

The fixed-update model predicts a linearly increasing amount of temporal recalibration as with increases in the asynchrony one is exposed to. However, empirical observations do not fully align with this model. The amount of audiovisual temporal recalibration as a function of adapted stimulus-onset asynchrony (SOA) exhibits two crucial characteristics: nonlinearity and asymmetry. The amount of recalibration is not proportional to the adapted SOA, but instead plateaus at SOAs of approximately 100–300 ms (Fujisaki et al., 2004; Vroomen et al., 2004). The amount of recalibration can also be asymmetrical: the magnitude of recalibration differs when the visual stimulus leads during the exposure phase compared to when the auditory stimulus leads (Fujisaki et al., 2004; O'Donohue et al., 2022; Van der Burg et al., 2013). These observations suggest that while the fixed-update model might capture the general purpose behind temporal recalibration, it falls short in fully capturing the nuanced ways in which the brain adjusts to varying SOAs. This prompts consideration of additional mechanisms to account for previously observed nonlinearity and asymmetry.

Notably, the fixed-update model overlooks the causal relationship between multimodal stimuli by implicitly assuming that they originate from a single source. However, that's not always the case in a multimodal environment. For instance, in a dubbed movie with a noticeable delay between the video and the audio, this delay can indicate that the sound is provided by a different actor, not the character on screen. To address this challenge, the brain must perform causal inference to determine whether multimodal signals come from a common source and should be integrated, or else kept separate. Indeed, numerous studies support the hypothesis that humans consider the causal structure of cross-modal stimuli when making perceptual decisions (Aller & Noppeney, 2019; Cao et al., 2019; Dokka et al., 2019; Körding et al., 2007; Locke & Landy, 2017; McGovern et al., 2016; Rohe & Noppeney, 2015; Samad et al., 2015; Sato et al., 2007; Wei & Körding, 2009; Wozny et al., 2010). Drawing on this framework, we formulate a causal-

inference model of temporal recalibration based on previous models that have successfully predicted visual-auditory (Hong et al., 2021; Sato et al., 2007) and visual-tactile (Badde, Navarro, & Landy, 2020) spatial recalibration.

Although models incorporating causal inference are promising in capturing the observed nonlinearities, they predict an identical amount of temporal recalibration for audiovisual stimulus pairs that have the same SOA but with opposite sign (i.e., lead vs. lag). This suggests that additional factors are required to explain the observed asymmetry. In previous studies, the asymmetry has been attributed to different factors, such as the physical and neural latency differences between sensory signals (O'Donohue et al., 2022; Van der Burg et al., 2013) or more frequent exposure to visual-lead events in natural environments (Fujisaki et al., 2004; Van der Burg et al., 2013). These factors can explain the audiovisual temporal bias most humans developed through early sensory experience (Badde, Ley, et al., 2020). Yet, this bias would again equally affect the amount of recalibration resulting from the same SOAs on either side of the observer's bias. In contrast to bias, sensory uncertainty has been shown to affect the degree of cross-modal recalibration in a complex fashion (Badde, Navarro, & Landy, 2020; Hong et al., 2021; van Beers et al., 2002). We hypothesize that different degrees of auditory and visual uncertainty play a critical role in the asymmetry of cross-modal temporal recalibration.

To examine the mechanism underlying cross-modal temporal recalibration, we used a classic three-phase recalibration paradigm, in which participants completed a pre-test, exposure, and post-test. We manipulated the adapter SOA (i.e., the audiovisual asynchrony presented in the exposure phase) across sessions, introducing SOAs up to 0.7 s of either auditory or visual lead. Before and after the exposure phase in each session, we measured participant's perception of audiovisual relative timing using a temporal-order-judgement (TOJ) task. To preview the empirical results, we confirmed the nonlinearity as well as idiosyncratic asymmetry of the recalibration effect. To scrutinize the factors that might drive these two main characteristics, we fitted four models to the data, using either causal inference or a fixed update. Despite previous empirical evidence challenging the fixed-update model, it doesn't mean we should discount its relevance without a statistical comparison to alternative models. The causal-inference and the fixed-update models were combined with either modality-specific or modality-independent uncertainty.

Model comparison revealed that causal inference combined with modality-specific uncertainty is essential to accurately capture the nonlinearity and idiosyncratic asymmetry of temporal recalibration. Our results indicate that human observers employ causal-inference-based percepts to recalibrate cross-modal temporal perception. This finding suggests that cross-modal temporal recalibration, typically considered an early-stage, low-level perceptual process, involves higher cognitive functions in the adjustment of perception.

2 Results

2.1 Behavioral results

We adopted a classical three-phase recalibration paradigm in which participants completed a pre-test, an exposure phase, and a post-test in each session. In pre- and post-tests, we measured participants' perception of audiovisual relative timing using a TOJ task: participants reported the perceived order ("visual first," "auditory first," or "simultaneous") of audiovisual stimulus pairs with varying SOAs (range: from -0.5 to 0.5 s with 15 levels; **Figure 1A**). In the exposure phase, participants were exposed to a series of audiovisual stimuli with a consistent SOA (250 trials; **Figure 1B**). To ensure that participants were attentive to the stimuli, they performed an oddball-detection task. Specifically, we inserted oddball stimuli with slightly greater intensity in either one or both modalities (5% of total trials independently sampled for each modality). Participants were instructed to respond whenever they detected such stimuli. The high d' of oddball-detection

performance (auditory $d' = 3.34 \pm 0.54$, visual $d' = 2.44 \pm 0.72$) showed that participants paid attention to both modalities (Figure S2). The post-test was almost identical to the pre-test, except that before every temporal-order judgment, there were three top-up exposure trials to maintain the recalibration effect. In total, participants completed nine sessions on separate days. The adapter SOA (range: -0.7 to 0.7 s) varied across but not within sessions.

We compared the temporal-order judgments between the pre- and post-tests to examine the amount of audiovisual temporal recalibration induced by the SOA of audiovisual stimuli during the exposure phase. Specifically, we fitted the data from the pre- and post-tests jointly assuming different points of subjective simultaneity (PSS) between the two tests while assuming fixed arrival-latency distributions and fixed response criteria (**Figure 2A**; see Supplement S1 for an alternative model assuming a shift in the response criteria due to recalibration). The amount of audiovisual temporal recalibration was defined as the difference between the two PSSs. At the group level, we observed a nonlinear pattern of recalibration as a function of the adapter SOA: the amount of recalibration in the direction of the adapter SOA first increased but then plateaued with increasing magnitude of the adapter SOA presented during the exposure phase (**Figure 2B**). Additionally, we observed an asymmetry between auditory-lead and visual-lead adapter SOAs in the magnitude of recalibration at the group level, with auditory-lead adapter SOAs inducing a greater amount of recalibration (**Figure 2B**; see Figure S6 for individual participants' data). To quantify this asymmetry for each participant, we calculated an asymmetry index, defined as the sum of the recalibration effects across all sessions (zero: no evidence for asymmetry; positive values: greater recalibration given visual-lead adapters; negative: greater recalibration given auditory-lead adapters). For each participant, we bootstrapped the temporal-order judgments to obtain a 95% confidence interval for the asymmetry index. All participants' confidence intervals excluded zero, suggesting that all of them showed audiovisual asymmetry in temporal recalibration (Figure S3).

2.2 Modeling results

In the following sections, we describe our models for cross-modal temporal recalibration by first laying out the general assumptions of these models, and then explaining the differences between them. Then, we provide a comparison in terms of model performance and illustrate how well the models capture the observed data by generating model predictions.

2.2.1 General model assumptions

We formulated four models of cross-modal temporal recalibration (**Figure 3**). These models share several common assumptions. First, when an auditory and a visual signal are presented, the corresponding neural signals arrive in the relevant brain areas with a variable latency due to internal and external noise. The probability distribution of arrival latency is an exponential distribution (García-Pérez & Alcalá-Quintana, 2012) (**Figure 3A**). A simple derivation shows that the resulting measurements of SOA follow a double-exponential distribution (**Figure 3B**). The mode reflects the physical SOA plus the participant's audiovisual temporal bias. The slopes of the distribution reflect the uncertainties of the arrival latency; the steeper the slope, the less variable the latency, and the less uncertainty a Bayesian observer would have in a single trial. Second, these models define temporal recalibration as accumulating updates of the audiovisual bias after each encounter with a SOA. The accumulated update of the audiovisual bias at the end of the exposure phase is then carried over to the post-test and persists throughout that phase. Lastly, the bias is assumed to be reset to the same initial value in the pre-test across all nine sessions, reflecting the stability of the audiovisual temporal bias across time (Badde, Ley, et al., 2020; Grabot & van Wassenhove, 2017).

Figure 1

Task timing. (A) Temporal-order-judgment task administered in the pre- and post-tests. In each trial, participants made a temporal-order judgment in response to an audiovisual stimulus pair with a varying stimulus-onset asynchrony (SOA). Negative values: auditory lead; positive values: visual lead. (B) Oddball-detection task used in the exposure phase and post-test top-up trials. Participants were repeatedly presented with an audiovisual stimulus pair with a SOA that was fixed within each session but varied across sessions. Occasionally, the intensity of either or both of the auditory and the visual stimuli was increased. Participants were instructed to press a key whenever such an oddball stimulus occurred.

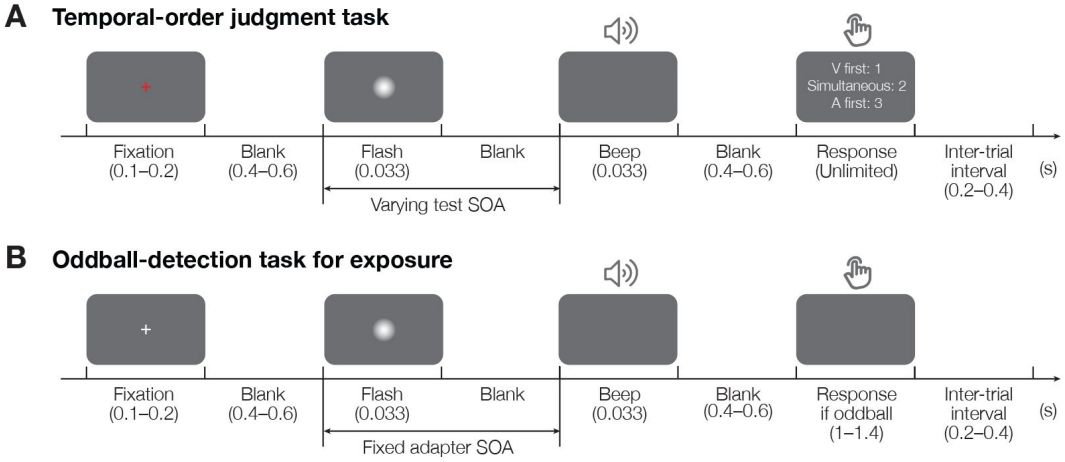
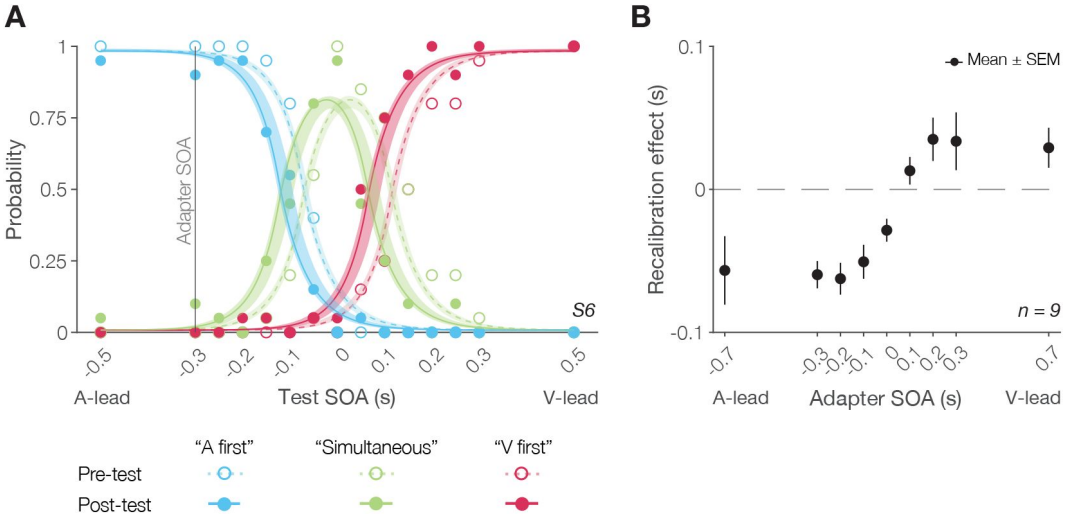


Figure 2

Behavioral results. (A) The probability of reporting that the auditory stimulus came first, the two arrived at the same time, or the visual stimulus came first as a function of SOA for a representative participant in a single session. The adapter SOA was -0.3 s for this session. Curves: best-fitting functions estimated jointly using the data from the pre-test (dashed) and post-test (solid). Shaded areas: 95% bootstrapped confidence intervals. (B) Mean recalibration effects (shifts in the point of subjective simultaneity from the pre- to the post-test phase) averaged across all participants as a function of adapter SOA. Error bars: ±SEM.



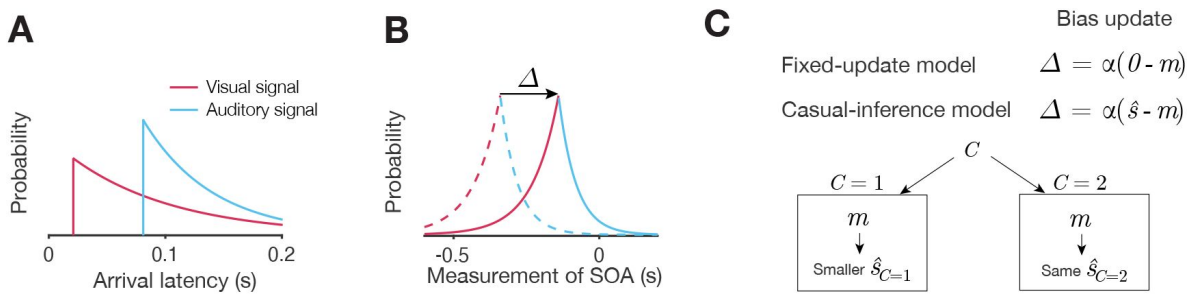


Figure 3

Illustration of the model for cross-modal temporal recalibration. (A) The probability density function of the arrival latency of the auditory and visual signals relative to the physical onset of each stimulus. (B) The resulting probability density function of the measured SOA, *m*, before (dashed) and after (solid) recalibration. The measurement distribution peaks at the physical SOA plus an audiovisual temporal bias. Temporal recalibration is modeled as cumulative changes in the audiovisual bias, Δ, across the exposure phase. (C) Two recalibration models. The fixed-update model updates the audiovisual bias so that subsequent measurements of the audiovisual SOA approach zero. The causal-inference model updates the audiovisual bias based on the perceived SOA, $\hat{s}$, i.e., taking different causal scenarios into account. The percept $\hat{s}$ is computed as a weighted average of estimates inferred from the scenarios of a common cause, $C = 1$, and separate causes, $C = 2$. α: learning rate. See text for details.

2.2.2 Models of cross-modal temporal recalibration

The four models we tested differed in the mechanism governing the updates of the audiovisual bias during the exposure phase as well as the modality specificity of the arrival latency uncertainty.

We formulated a Bayesian causal-inference model (Körding et al., 2007; McGovern et al., 2016; Sato et al., 2007) to describe the recalibration of the relative timing between cross-modal signals. When an observer is presented with an adapter SOA during the exposure phase, they infer the causal relationship between the auditory and visual stimulus. Specifically, the observer computes two intermediate estimates of the SOA, one for each causal scenario (**Figure 3C**). In the common-cause scenario, the estimated SOA of the stimuli is smaller than the measurement as it is combined with a prior distribution over SOA that reflects synchrony. In the separate-causes scenario, the estimated SOA is approximately equal to the measurement. The two estimates are then averaged with each one weighted by the inferred probability of the corresponding causal scenario. The audiovisual bias is then updated to reduce the difference between the measurement and the combined estimate of SOA. In other words, causal inference regulates the recalibration process by shifting the measured SOA to more closely match the percept, which in turn is computed based on the inferred causal structure.

We also considered a fixed-update model. The major distinction between the causal-inference and the fixed-update model is that, according to the latter, the measured SOA is shifted toward zero rather than toward the inferred SOA. Essentially, whenever the observer detects a SOA, they recalibrate by shifting the audiovisual bias in the opposite direction so that the measured SOA will be closer to zero.

We additionally varied a second model element: we assumed either modality-specific or modality-independent uncertainty of arrival latency. The auditory system typically has higher temporal precision than the visual system. Hence, the arrival latency of visual signals can be more variable than auditory-signal latency, resulting in an asymmetrical probability density of measured SOA (m). A Bayesian observer will take this modality-specific sensory uncertainty into account to derive an estimate of SOA ($\hat{s}$). However, temporal precision might not be due to the variability of arrival latency. The auditory and visual systems might share a common, modality-independent timing mechanism (Stauffer et al., 2012), predicting modality-independent uncertainty.

2.2.3 Model comparison

We fitted four models to each participant's data. Each model was constrained jointly by the temporal-order judgments from the pre- and post-tests of all nine sessions. To quantify model performance, we computed the Akaike information criterion (AIC) for each model and each participant (Akaike, 1998). The model with the lowest AIC value was considered the best-fitting model. For all participants, the causal-inference model with modality-specific uncertainty outperformed the other three models. We then computed the AIC values of the other models relative to the best-fitting model, ΔAIC, with higher Δ_{AIC} values indicating stronger evidence for the best-fitting model. The results of model comparison revealed robust evidence for the causal-inference model with modality-specific uncertainty (Δ_{AIC} = 55.68 ± 21.45 for the fixed-update model with modality-specific uncertainty; Δ_{AIC} = 48.71 ± 18.51 for the fixed-update model with modality-independent uncertainty; Δ_{AIC} = 12 ± 5.94 for the causal-inference model with modality-independent uncertainty).

2.2.4 Model prediction

We predicted the recalibration effect per adapter SOA using the estimated parameters based on each of the four models. The nonlinearity in audiovisual temporal recalibration was only captured by models that rely on causal inference during the exposure phase (**Figure 4A**; see Figure S5 for other variants of the causal-inference model; see Figures S6 and S7 for model predictions for individual participants' recalibration effects and TOJ responses). On the other hand, the models that assume a fixed update based on the measured SOA were unable to capture the data, as they predict a linear increase of recalibration with greater adapter SOA. We derived the asymmetry index (i.e., the recalibration effect summed across sessions) for the predictions of each model and compared these indices with those computed directly from the data. To capture participants' idiosyncratic asymmetry in temporal recalibration, the model not only requires modality-specific uncertainty of arrival latency, it also needs to account for causal inference during the exposure phase (**Figure 4B**).

2.2.5 Model simulation

Simulations with the best-fitting model revealed key factors that determine the degree of non-linearity and asymmetry of cross-modal temporal recalibration to different adapter SOAs. The belief that the auditory and visual stimuli share a common cause plays a crucial role in adjudicating between these two causal scenarios (**Figure 5A**). When the observer infers that the audiovisual stimuli share the same cause, they recalibrate by a proportion of the perceived asynchrony no matter how large the measured asynchrony is, identical to the fixed-update model. On the contrary, when the observer infers that the audiovisual stimuli have separate causes, they treat the audiovisual stimuli as independent of each other and do not recalibrate. Estimates of the common-cause prior for all participants range between these two extremes. Thus, all observers weighted the estimates from these two scenarios based on the scenarios' probability, resulting in the nonlinear pattern of recalibration (see Table S1 for parameter estimates for individual participants).

Differences in arrival-time uncertainty between audition and vision result in an asymmetry of audiovisual temporal recalibration across adapter SOAs (**Figure 5B**). The amount of recalibration is attenuated when the modality with less uncertainty lags during the exposure phase. When the lagging stimulus is less uncertain, the perceptual system is more likely to attribute the SOA to separate causes and thus recalibrate less. In addition, the initial audiovisual bias does not affect asymmetry, but shifts the recalibration function horizontally and determines the SOA for which no recalibration occurs (Figure S8).

3 Discussion

In this study, we examined audiovisual temporal recalibration by repeatedly exposing participants to various stimulus-onset asynchronies and measured perceived audiovisual relative timing before and after exposure. To further understand the mechanisms underlying audiovisual temporal recalibration, we assessed the efficacy of different models of the recalibration process in predicting the amount of recalibration as a function of the audiovisual asynchrony to which one is exposed. Our findings indicate that a Bayesian causal-inference model with modality-specific uncertainty best captured the two key features of cross-modal temporal recalibration: the non-linear increase of recalibration magnitude with increasing adapted audiovisual asynchrony, and the asymmetrical recalibration magnitude between auditory- and visual-lead adapters with the same absolute asynchrony. Our results indicate that human observers employ causal-inference-based percepts to recalibrate cross-modal temporal perception.

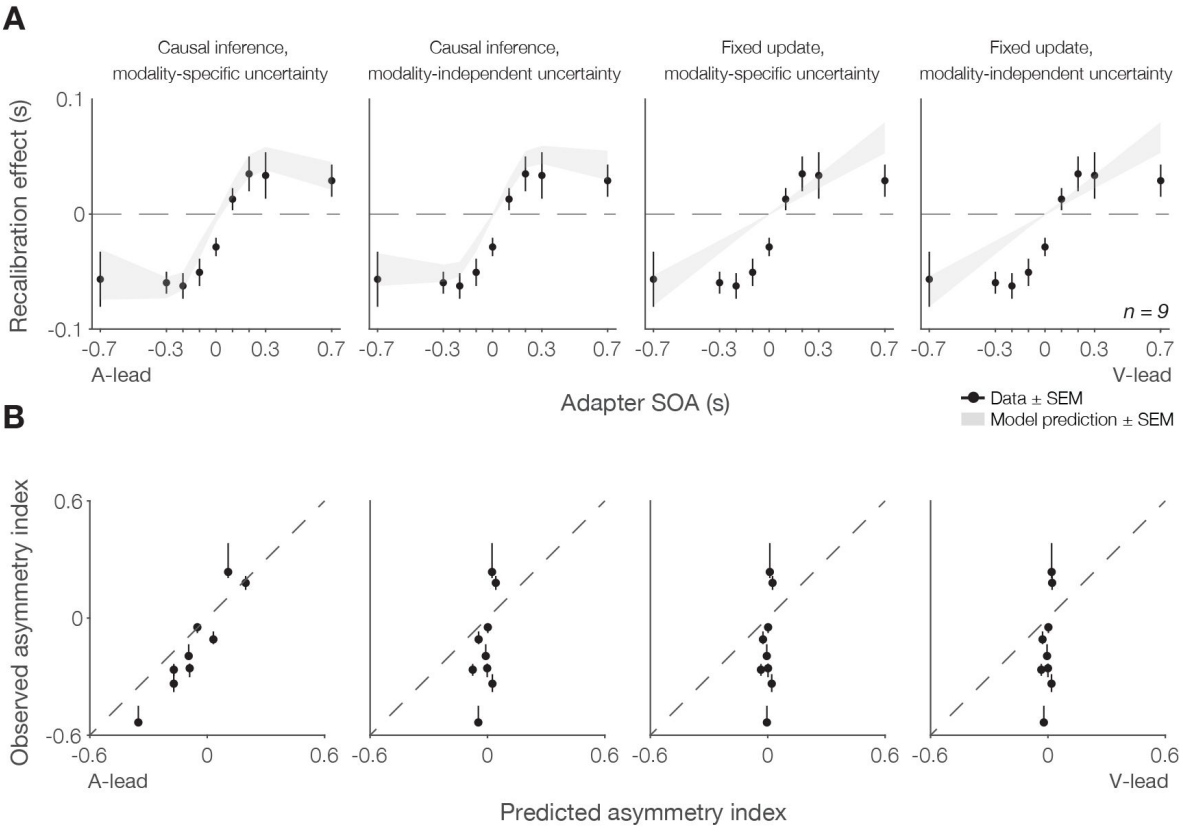


Figure 4

Model predictions. (A) Data and model predictions of recalibration as a function of adapter SOA. (B) Model prediction of the asymmetry index, the summed recalibration effect across adapter SOA. Dots: individual participants. Error bars: 68% bootstrapped confidence intervals. Identity line: perfect model prediction.

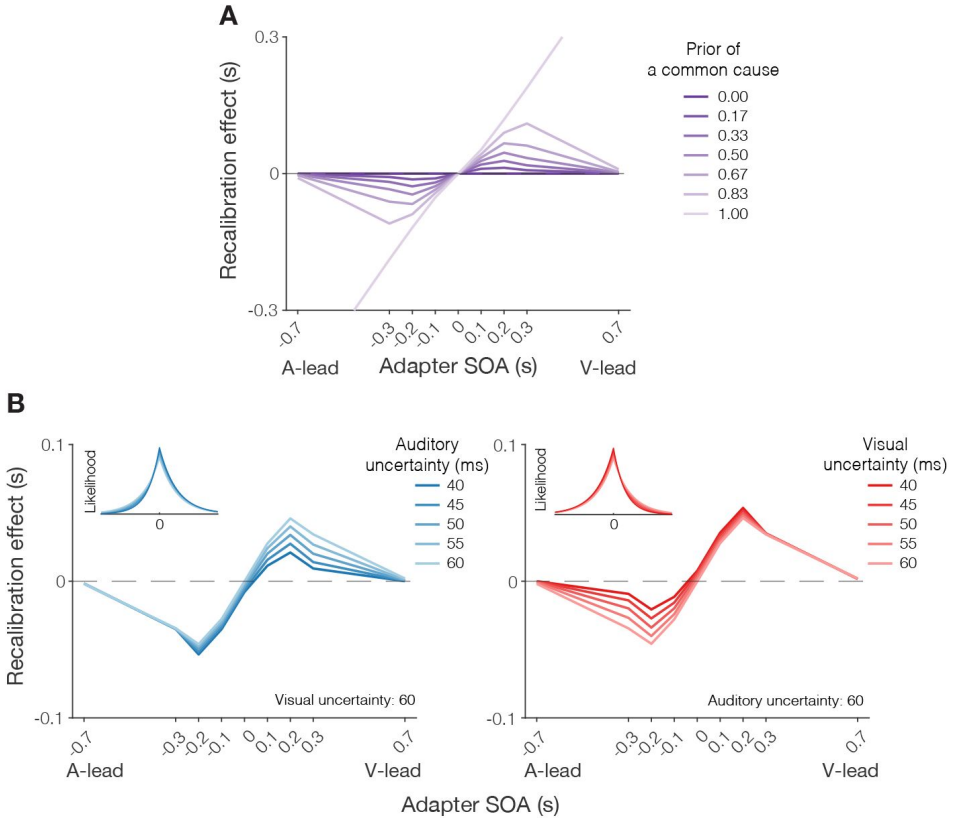


Figure 5

Model simulations. (A) Effect of the prior probability of a common cause on cross-modal temporal recalibration. (B) Asymmetry due to differing uncertainty of auditory vs. visual arrival latency. Left panel: When auditory uncertainty is smaller than visual uncertainty, reducing auditory uncertainty leads to less recalibration in response to visual-lead adapter SOA. Right panel: when visual uncertainty is smaller than auditory uncertainty, the opposite effect results. Top-left insets: corresponding SOA likelihood functions for a measured SOA of zero as auditory or visual uncertainty is varied.

In cross-modal recalibration, causal inference effectively serves as a credit-assignment mechanism, evaluating to what extent the source of discrepancy is external (i.e., the stimuli from the two modalities occurred at different times in the world) or internal (i.e., the measurement of asynchrony resulted from internal miscalibration). The perceptual system should correct for errors if they are due to misalignment between the senses. It shouldn't recalibrate if two independent events, such as your TV screen and the neighbors' conversation, exhibit audiovisual asynchrony. The same principle also applies to other cross-modal domains. The relevance of causal inference extends beyond temporal recalibration, influencing cross-modal spatial recalibration (Badde, Navarro, & Landy, 2020; Hong et al., 2021; Sato et al., 2007; Wozny & Shams, 2011a, 2011b). Similarly, in sensorimotor adaptation, humans correct for motor errors that are more likely due to the motor system, but not due to the environment (Berniker & Kording, 2008; Wei & Körding, 2009).

Previous investigations into the mechanisms behind audiovisual temporal recalibration have proposed various models. These models describe recalibration as a selective reduction of response gain of the adapted asynchrony in a population code (Roach et al., 2011), a shift of latency or response criteria (Yarrow et al., 2015), changes in temporal discriminability (Roseboom et al., 2015), or the update of prior and likelihood function (Sato & Aihara, 2011). However, a common feature of the experimental methods in these studies is to examine the recalibration process within a relatively narrow range of adapted audiovisual asynchrony. This is based on the assumption that the audiovisual stimuli are perceived as originating from the same source, which holds for small asynchronies. Our model seeks to go beyond this limitation by incorporating causal inference, which extends the model applicability across a wider range of audiovisual asynchrony.

In addition to the nonlinear pattern of temporal recalibration, our results revealed significant asymmetry in how much participants recalibrated to visual- vs. auditory-lead stimuli. The majority of our participants showed larger recalibration effects in response to auditory-than visual-lead asynchrony, in line with previous studies (O'Donohue et al., 2022). Simulation results supported the idea that this asymmetry could be due to less uncertainty in auditory arrival latency, in line with psychophysical studies (reviewed by Stauffer et al., 2012) that found audition has better temporal sensitivity than vision. However, our findings also highlighted individual differences: a few participants showed the opposite pattern, which was also revealed before (Fujisaki et al., 2004). We speculate that in certain conditions, visual arrival-latency might be less variable than auditory latency if the auditory signal is influenced by environmental factors such as echoes. Accordingly, our model explains how temporal uncertainty, based on the precision of the perceptual system and temporal variability of the physical stimulus, can lead to different directions of asymmetry in audiovisual temporal recalibration.

The principle of causal inference in audiovisual temporal recalibration is likely to extend to rapid cross-modal temporal recalibration, which occurs following the exposure to a single and brief audiovisual asynchrony (Van der Burg et al., 2013). However, it is an open question whether modality-specific uncertainty can explain the asymmetry of rapid cross-modal temporal recalibration. The pattern of asymmetry for rapid temporal recalibration differs from that of cumulative recalibration; in rapid recalibration, recalibration magnitude and the range in which recalibration occurs is larger when vision leads than when audition leads (Van der Burg et al., 2013, 2015). Such findings suggest that the mechanisms behind rapid and cumulative temporal recalibration may differ fundamentally. Supporting this, recent neuroimaging research has revealed distinct underlying neurophysiological processes. Cumulative temporal recalibration induces gradual phase shifts of entrained neural oscillations in the auditory cortex (Kösem et al., 2014), whereas rapid recalibration relies on phase-frequency coupling that happens at a faster time scale (Lennert et al., 2021).

In sum, we found that causal inference with modality-specific uncertainty modulates audiovisual temporal recalibration. This finding suggests that cross-modal temporal recalibration is more complex than a compensatory mechanism for maintaining accuracy or consistency. It relies on causal inference that considers both the sensory and causal uncertainty of multisensory inputs. Cross-modal temporal recalibration is typically viewed as an early-stage, low-level perceptual process. Our findings refine this view, suggesting that it is deeply intertwined with higher cognitive functions.

4 Methods

4.1 Participants

Ten students from New York University (three males; age: 24.4 ± 1.77; all right-handed) participated in the experiment. They all reported normal or corrected-to-normal vision. All participants provided informed written consent before the experiment and received $15/hr as monetary compensation. The study was conducted in accordance with the guidelines laid down in the Declaration of Helsinki and approved by the New York University institutional review board. Data of one of the participants was identified as an outlier and therefore excluded from further data analysis (Figure S4).

4.2 Apparatus and stimuli

Participants completed the experiments in a dark and semi sound-attenuated room. They were seated 1 m from an acoustically transparent, white screen (1.36 × 1.02 m, 68 × 52° visual angle) and placed their head on a chin rest. An LCD projector (Hitachi CP-X3010N, 1024 × 768 pixels, 60 Hz) was mounted above and behind participants to project visual stimuli on the screen. The visual stimulus was a high-contrast (36.1 cd/m^2) Gaussian blob (SD: 3.6°) on a gray background (10.2 cd/m^2) projected onto the screen. The auditory stimulus was a 500 Hz beep (50 dB SPL) played by a loudspeaker behind and located at the center of the screen. The visual and auditory stimulus durations were 33.33 ms. We adjusted the timing of audiovisual stimulus presentations and verified the timing using an oscilloscope (PICOSCOPE 2204A).

4.3 Procedure

The experiment consisted of nine sessions, which took place on nine separate days. In each session, participants completed a pre-test, an exposure phase, and a post-test in sequence. The adapter SOA was fixed within a session, but varied across sessions (±700, ±300, ±200, ±100, 0 ms). The intensities of the oddball stimuli were determined prior to the experiment for each participant using an intensity-discrimination task to equate the difficulty of detecting oddball stimuli between participants and across modalities.

4.3.1 Pre-test phase

Participants completed a TOJ task during the pre-test phase. Each trial started with the display of a fixation cross (0.1–0.2 s, uniform distribution), followed by a blank screen (0.4–0.6 s, uniform distribution). Then, an auditory and a visual stimulus (0.033 s) were presented with a variable SOA. There were a total of 15 possible test SOAs (from -0.5 to 0.5 s in steps of 0.05 s), with positive values representing visual lead and negative values representing auditory lead. Following stimulus presentation there was another blank screen (0.4–0.6 s, uniform distribution), and then a response probe appeared on the screen. Participants indicated by button press whether the auditory stimulus occurred before the visual stimulus, occurred after, or the two were simultaneous. There was no time limit for the response, and response feedback was not provided.

The inter-trial interval (ITI) was 0.2–0.4 s. Each test SOA was presented 20 times in pseudo-randomized order, resulting in 300 trials in total, divided into five blocks. Participants usually took around 15 minutes to finish the pre-test phase.

4.3.2 Exposure phase

Participants completed an oddball-detection task during the exposure phase. In each trial, participants were presented with an audiovisual stimulus pair with a fixed SOA (adapter SOA). In 10% of trials, the intensity of either the visual or the auditory component (or both) was greater than in the other trials. Participants were instructed to press a button as soon as possible when there was an auditory oddball, a visual oddball, or both stimuli were oddballs. The task timing was almost identical to the TOJ task, except that there was a response time limit of 1.4 s. The visual and auditory oddball stimuli were presented to participants prior to the exposure phase and they practiced as much as they needed to familiarize themselves with the task. There were a total of 250 trials, divided into five blocks. At the end of each block, we presented a performance summary with the hit and false alarm rates for each modality. Participants usually took 15 minutes to complete the exposure phase.

4.3.3 Post-test phase

Participants completed the TOJ task as well as the oddball-detection task during the post-test phase. Specifically, each temporal-order judgment was preceded by three top-up (oddball-detection) trials. The adapter SOA in the top-up trials was the same as that in the exposure phase to prevent dissipation of temporal recalibration (Machulla et al., 2012). To facilitate task switching, the ITI between the last top-up trial and the following TOJ trial was longer (with the additional time jittered around 1 s). Additionally, the fixation cross became red to signal the start of a TOJ trial. As in the pre-test phase, there were 300 TOJ trials (15 test SOAs × 20 repetitions) with the addition of 900 top-up trials, grouped into six blocks. At the end of each block, we provided a summary of the oddball-detection performance. Participants usually took around 1 hour to complete the post-test phase.

4.3.4 Intensity-discrimination task

This task was conducted to estimate the just-noticeable-difference (JND) in intensity for a standard visual stimulus with a luminance of 36.1 cd/m^2, and a standard auditory stimulus with a volume of 40 dB SPL. The task was two-interval, forced-choice. The trial started with a fixation (0.1–0.2 s) and a blank screen (0.4–0.6 s). Participants were presented with a standard stimulus in one randomly selected interval (0.4–0.6 s) and a comparison stimulus in the other interval (0.4–0.6 s), temporally separated by an inter-stimulus interval (0.6–0.8 s). They indicated which interval contained the brighter/louder stimulus without time constraint. Seven test stimulus levels (luminance range: 5%–195%; volume range: 50%–150% of the standard) were repeated 20 times, resulting in 140 trials for each task. We fit a Gaussian cumulative distribution function to these data and defined the JND as the intensity difference for which the test stimulus was chosen 90% of the time as more intense than the standard. An oddball was defined as an auditory or visual stimulus with an intensity 1 JND above the standard intensity.

4.4 Modeling

In this section, we use the best-fitting model that combines causal inference with modality-specific uncertainty as a template, and then describe how the alternative models differ from this. We start by describing how the arrival latencies of auditory and visual stimuli lead to noisy internal measurements of audiovisual SOA, followed by how we modeled the process of audiovisual temporal recalibration. Then, we provide a formalization of the TOJ task administered in the pre- and the post-test phases, data from which were used to constrain the model parameters. Finally, we describe how the models were fit to the data.

4.4.1 Measurements of audiovisual stimulus-onset-asynchrony

When an audiovisual stimulus pair with a stimulus onset asynchrony, $s = t_A - t_V$, is presented, it triggers auditory and visual signals that are registered with different latency in the region of cortex where audiovisual comparisons are made. This leads to two internal measurements in an observer's brain. As in previous work (Garcí a-Pé rez & Alcalá -Quintana, 2012), we model the probability of the latency of auditory and visual signals across repetitions, relative to the physical onset of each stimulus, as shifted exponential distributions (**Figure 3A**). These distributions may be shifted relative to the physical stimulus onset due to internal signal delays (denoted β_V and β_A). The arrival latency of the auditory signal relative to onset t_A is the sum of the fixed delay, β_A, and an additional delay that is exponentially distributed with time constant τ_A, and similarly for the visual arrival latency (with delay β_V and time constant τ_V).

The measured SOA of the audiovisual stimulus pair is modeled as the difference in the arrival latency of both stimuli. Thus, the measured audiovisual SOA m includes the physical SOA s, the fixed latency (i.e., the difference between the auditory and visual fixed latency) $\beta = \beta_A - \beta_V$, and a stochastic component. Given that both latency distributions are shifted exponential distributions, a noisy sensory measurement of SOA m given a physical SOA s has a probability density that is an asymmetric double-exponential (**Figure 6A**):

$$P(m|s,\beta) = \begin{cases} \frac{1}{\tau_A + \tau_V} \exp\left[\tau_V^{-1}(m - (s+\beta))\right], & \text{if } m \le s+\beta, \\ \frac{1}{\tau_A + \tau_V} \exp\left[-\tau_A^{-1}(m - (s+\beta))\right], & \text{if } m > s+\beta. \end{cases} \quad (1)$$

The mode of this measurement distribution is the physical SOA plus the fixed latency $s + \beta$. A negative value of β indicates faster auditory processing. The left and right spread of this measurement distribution depends on the uncertainty of the visual latency τ_V and auditory latency τ_A, respectively.

4.4.2 The perceptual inference process

To infer the stimulus SOA s from the measurement m, the ideal observer initially computes the posterior distribution of the SOA s by multiplying the likelihood function and the prior for two causal scenarios. The auditory and visual stimuli can arise from a single cause ($C = 1$) or two independent causes ($C = 2$).

The ideal observer holds two prior distributions of audiovisual SOA, one for each causal scenario. In the case of a common cause ($C = 1$), the prior distribution of the SOA between sound and light is a narrow Gaussian distribution (McGovern et al., 2016),

$$P(s|C=1) = \mathcal{N}(0, \sigma^2_{C=1}). \quad (2)$$

When there are two separate causes ($C = 2$), the prior distribution of the audiovisual SOA is a broad Gaussian distribution (McGovern et al., 2016), assigning almost equal probability to each audiovisual SOA

$$P(s|C=2) = \mathcal{N}(0, \sigma^2_{C=2}). \quad (3)$$

The observer obtains intermediate estimates of the stimulus SOA by combining the measured SOA with the prior over SOA corresponding to the two causal scenarios, $\hat{s}_{C=1}$ and $\hat{s}_{C=2}$. In this model, we assume that this observer doesn't have access to, or chooses to ignore, the current temporal bias β.

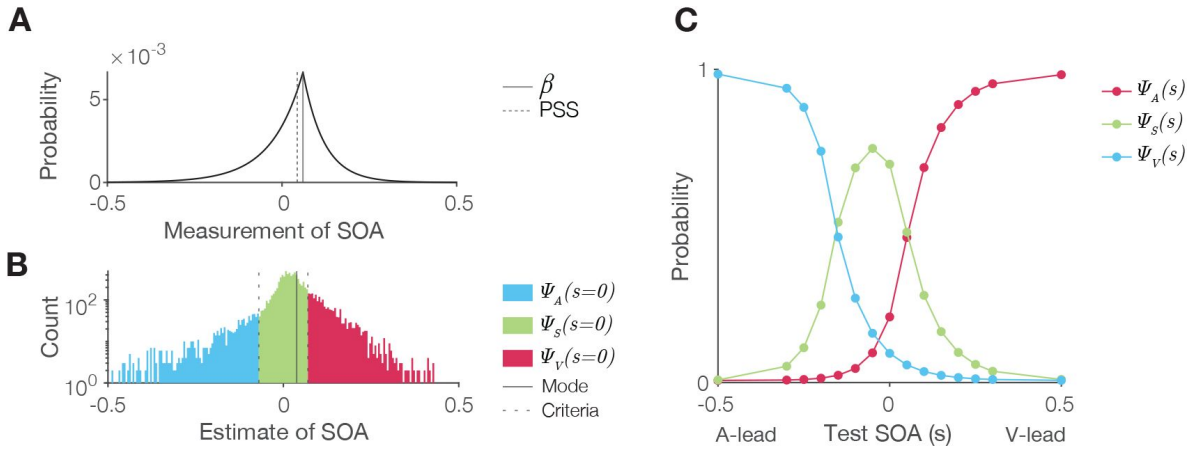


Figure 6

Causal-inference model of the temporal-order-judgment task. (A) An example measurement distribution for a SOA of zero. The measurement distribution peaks at the audiovisual bias β, whose left and right slopes reflect the visual and auditory uncertainty, respectively. The point of subjective simultaneity is marked by the dashed line. (B) Simulated estimate distribution for a SOA of zero. The dashed lines represent the criteria placed symmetrically around zero, forming a temporal window of SOA estimates treated as simultaneous. The areas under the estimate distribution partitioned by the criteria indicate the probabilities of the three possible responses for a stimulus pair with a SOA of zero. (C) Simulated psychometric function computed by repeatedly calculating the probability of each possible response for all SOAs.

The likelihood functions under the two causal scenarios are identical:

$$\begin{aligned}\mathcal{L}(s|m) &= P(m|s) \\ &= \begin{cases} \frac{1}{\tau_A+\tau_V}\exp\left[\tau_A^{-1}(s-m)\right], & \text{if } s < m, \\ \frac{1}{\tau_A+\tau_V}\exp\left[-\tau_V^{-1}(s-m)\right], & \text{if } s \geq m. \end{cases}\end{aligned} \tag{4}$$

where the left and right spreads depend on auditory and visual uncertainties of the arrival latency. Because the likelihood function is non-Gaussian, there is no closed form for the intermediate estimate. We computed the posterior numerically and used a maximum-a-posteriori (MAP) estimator, i.e., $\hat{s}$ was the mode of the posterior over stimulus SOA in each scenario.

The final estimate of the stimulus SOA $\hat{s}$ depends on the posterior probability of each causal scenario. By Bayes Rule, the posterior probability that an audiovisual stimulus pair with the measured SOA shares a common cause is

$$P(C=1|m) = \frac{P(m|C=1)P(C=1)}{P(m|C=1)P(C=1)+P(m|C=2)(1-P(C=1))}. \tag{5}$$

The likelihood of a common source/separate sources for a fixed SOA measurement is calculated by numerically integrating the protoposterior (i.e., the unnormalized posterior),

$$\begin{aligned} P(m|C=1) &= \int P(m|s)P(s|C=1)ds, \\ P(m|C=2) &= \int P(m|s)P(s|C=2)ds. \end{aligned} \tag{6}$$

The posterior probability of a common cause additionally depends on the observer's prior belief of a common cause for the auditory and visual stimuli, $P(C=1) = p_{\text{common}}$.

The final estimate of SOA is derived by model averaging, so that the final estimate is the average of the scenario-specific SOA estimates above weighted by the posterior probability of the corresponding causal scenario,

$$\hat{s} = P(C=1|m)\hat{s}_{C=1} + (1-P(C=1|m))\hat{s}_{C=2}\,. \tag{7}$$

4.4.3 Formalization of recalibration in the exposure phase

We model the recalibration process as a shift of the audiovisual fixed latency β, audiovisual temporal bias, after encountering an audiovisual stimulus pair (**Figure 3B**). The internal value of β reflects the observed point of subjective simultaneity (PSS), which is the stimulus SOA s that leads to a median measurement of audiovisual synchrony equal to $m = 0$. That is, it is the value of s such that $P(m < 0 \,|\, s = \text{PSS}, \beta) = 0.5$. A simple derivation yields

$$\text{PSS} = \begin{cases} -\beta - \tau_V \ln\left(\frac{\tau_A+\tau_V}{2\tau_V}\right), \text{ if } \tau_A \leq \tau_V \\ -\beta + \tau_A \ln\left(\frac{\tau_A+\tau_V}{2\tau_A}\right), \text{ if } \tau_A > \tau_V\,. \end{cases} \tag{8}$$

The shift of the audiovisual bias β also moves the measurement distribution. We assume the exponential time constants (τ_A, τ_V) remain unchanged across phases and sessions.

At the end of every exposure trial i, a discrepancy between the measured SOA, m_i and the final estimate of the stimulus SOA $\hat{s}_i$ signals the need for recalibration. In each session, we assume the participant arrives with a default bias β. We define $\Delta_{\beta,i}$ as the cumulative shift of audiovisual bias

after exposure trial i,

$$\Delta_{\beta,i+1} = \Delta_{\beta,i} + \alpha(\hat{s}_i - m_i), \tag{9}$$

where α is the learning rate. At the end of the exposure phase, the predicted audiovisual bias is thus shifted by the accumulated shifts across the exposure phase, that is, $\beta_{\text{post}} = \beta + \Delta_{\beta,250}$.

4.4.4 Formalization of the temporal-order-judgement task

In the TOJ task administered in the pre- and post-test phases, the observer makes a perceptual judgment by comparing the final estimate of stimulus SOA $\hat{s}$ to two internal criteria (Cary et al., 2024; García-Pérez & Alcalá-Quintana, 2012). We assume that the observer has a symmetric pair of criteria, $\pm c$, centered on the stimulus SOA corresponding to perceptual simultaneity ($\hat{s} = 0$). In addition, the observer may lapse or make an error when responding. The probabilities of reporting visual lead, Ψ_V, auditory lead, Ψ_A or that the two stimuli were simultaneous, Ψ_S, are thus

$$\begin{aligned}
\Psi_V(s) &= \frac{\lambda}{3} + (1-\lambda)\tilde{P}(\hat{s} > c|\beta), \\
\Psi_A(s) &= \frac{\lambda}{3} + (1-\lambda)\tilde{P}(\hat{s} < -c|\beta) \text{ and} \\
\Psi_S(s) &= 1 - \Psi_V(s) - \Psi_A(s).
\end{aligned} \tag{10}$$

where λ is the lapse rate. **Figure 6C** shows an example of the resulting psychometric functions.

The probability distribution of causal-inference-based stimulus SOA estimates $P(\hat{s}|s)$ has no closed form and can only be simulated. For each simulation we sampled 10,000 SOA measurements from the corresponding double-exponential probability distribution (**Figure 6A**). For each sampled measurement, we simulated the process by which the observer carries out causal inference and produced an estimate of the stimulus SOA (fixing the values of a few additional causal-inference model parameters). This process resulted in a Monte-Carlo approximation of the probability distribution of the causal-inference-based stimulus SOA estimates (**Figure 6B**).

4.4.5 Alternative models

In the fixed-update model, observers measure the audiovisual SOA s by comparing the arrival latency of the auditory and visual signals (**Eq. 1**). They do not perform causal inference to estimate the SOA. Instead, the measured SOA is shifted toward zero by recalibrating the audiovisual bias β. Hence, the update of the audiovisual bias in trial i is defined by

$$\Delta_{\beta,i+1} = \Delta_{\beta,i} - \alpha m_i. \tag{11}$$

The update of audiovisual bias is accumulated across the exposure phase, $\beta_{\text{post}} = \beta + \Delta_{\beta,250}$. In TOJ tasks, observers make the temporal-order decision by applying the criteria to the measurement of SOA m (see psychometric functions in Supplement Eq. S1).

In models with modal-independent uncertainty, $\tau_A = \tau_V$, resulting in a symmetrical measurement distribution (**Eq. 1**).

4.4.6 Model fitting

Model log-likelihood

The model was fitted by maximizing likelihood. We fit the model to the TOJ data collected during the pre- and post-test phases of all sessions together. We did not collect temporal-order judgments in the exposure phase. However, to model the post-test data, we needed to estimate the distribution of shifts of audiovisual bias resulting from the exposure phase ($\Delta_{\beta,250}$). We did this using Monte Carlo simulation of the 250 exposure trials to estimate the probability distribution of the cumulative shifts.

The set of model parameters Θ is listed in **Table 1**. There are I sessions, each including K trials in the pre-test phase and K trials in the post-test phase. We denote the full dataset of pre-test data as X_{pre} and for the post-test data as X_{post}. On a given trial, the observer responds either auditory-first (A), visual-first (V) or simultaneous (S). We denote a single response using indicator variables that are equal to 1 if that was the response in that trial and 0 otherwise. These variables for trial k in session i are $r^{A}_{\text{pre},ik}$, $r^{V}_{\text{pre},ik}$ and $r^{S}_{\text{pre},ik}$ for the pre-test trials, and $r^{A}_{\text{post},ik}$, etc., for the post-test trials. The log-likelihood of all pre-test responses X_{pre} given the model parameters given is

$$\begin{aligned}\log p(X_{\text{pre}}|M,\Theta) = \sum_{i=1}^{I}\sum_{k=1}^{K}\Big(& r^{A}_{\text{pre},ik}\log\Psi_{A,\text{pre}}(s_{ik}) + \\ & r^{V}_{\text{pre},ik}\log\Psi_{V,\text{pre}}(s_{ik}) + r^{S}_{\text{pre},ik}\log\Psi_{S,\text{pre}}(s_{ik})\Big).\end{aligned} \tag{12}$$

The psychometric functions for the pre-test (e.g., $\Psi_{A,\text{pre}}$) are defined in **Eq. 11**, and are the same across all sessions as we assumed that the audiovisual delay β was the same before recalibration in every session.

The log-likelihood of responses in the post-test depends on the audiovisual bias after re-calibration $\beta_{\text{post}} = \beta + \Delta_{\beta,250,i}$ for session i. To determine the log-likelihood of the post-test data requires us to integrate out the unknown value of the cumulative shift $\Delta_{\beta,250,i}$. We approximated this integral in two steps based on our previous work (Hong et al., 2021). First, we simulated the 250 exposure-phase trials 1,000 times for a given set of parameters Θ and session i. This resulted in 1,000 values of $\Delta_{\beta,250,i}$. The distribution of these values was well fit by a Gaussian whose parameters were determined by the empirical mean and standard deviation of the sample distribution, resulting in the distribution $\tilde{P}(\Delta_{\beta,250,i}|M,\Theta)$. Second, we approximated the integral of the log-likelihood of the data over possible values of $\Delta_{\beta,250,i}$ by numerical integration. We discretized the approximated distribution $\tilde{P}(\Delta_{\beta,250,i}|M,\Theta)$ into 100 equally spaced bins centered on values $\Delta_{\beta,250,i}(n)$ (n = 1, ..., 100). The range of the bins was triple the range of the values from the Monte Carlo sample, so that the lower bound was $lb_{\Delta_{\beta,250,i}} = \Delta_{\beta,250,i,\min} - (\Delta_{\beta,250,i,\max} - \Delta_{\beta,250,i,\min})$ and the upper bound was $ub_{\Delta_{\beta,250,i}} = \Delta_{\beta,250,i,\max} + (\Delta_{\beta,250,i,\max} - \Delta_{\beta,250,i,\min})$.

The log-likelihood of the post-test data is approximated as

$$\begin{aligned}\log p(X_{\text{post}}|M,\Theta) &= \sum_{i=1}^{I}\log\left(\int P(X_{\text{post}}|\Delta_{\beta,250,i},M,\Theta)P(\Delta_{\beta,250,i}|M,\Theta)d\Delta_{\beta,250,i}\right) \\ &\approx \sum_{i=1}^{I}\log\left(\int_{lb_{\Delta_{\beta,250,i}}}^{ub_{\Delta_{\beta,250,i}}} P(X_{\text{post}}|\Delta_{\beta,250,i},M,\Theta)\times \right. \\ &\quad \left.\tilde{P}(\Delta_{\beta,250,i}|M,\Theta)d\Delta_{\beta,250,i}\right) \\ &\approx \sum_{i=1}^{I}\log\left(\frac{ub_{\Delta_{\beta,250,i}} - lb_{\Delta_{\beta,250,i}}}{100}\sum_{n=1}^{100} P(X_{\text{post}}|\Delta_{\beta,250,i}(n),M,\Theta)\times \right. \\ &\quad \left.\tilde{P}(\Delta_{\beta,250,i}(n)|M,\Theta)\right),\end{aligned} \tag{13}$$

Notation	Specification	Temporal-order-judgement task	Recalibration in the exposure phase
β	The fixed relative delay between visual and auditory processing, i.e., the audio-visual temporal bias prior to the exposure phase	✓	✓
τ_A	Auditory uncertainty, the exponential time constant of the auditory latency distribution	✓	✓
τ_V	Visual uncertainty, the exponential time constant of the visual latency distribution	✓	✓
$\sigma_{C=1}$	The width of the Gaussian prior for the common-cause scenario	✓	✓
$\sigma_{C=2}$	The width of the Gaussian prior for the separate-causes scenario	✓	✓
p_{common}	The prior probability of a common cause	✓	✓
c	Simultaneity criterion	✓	
λ	Lapse rate	✓	
α	Learning rate for shifting audiovisual temporal bias		✓

Table 1

Model parameters. Check marks signify that the parameter is used for determining the likelihood of the data from the temporal-order judgment task in the pre- and post-test phase and/or for the Monte Carlo simulation of recalibration in the exposure phase.

where

$$P(X_{\text{post}}|\Delta_{\beta,250,i}(n),M,\Theta)=\prod_{k=1}^{K}\left(\Psi_{A,\text{post},in}(s_{ik})^{r^{A}_{\text{post},ik}}\times \Psi_{V,\text{post},in}(s_{ik})^{r^{V}_{\text{post},ik}}\Psi_{S,\text{post},in}(s_{ik})^{r^{S}_{\text{post},ik}}\right). \quad (14)$$

The psychometric functions in the post-test (e.g., $\Psi_{A,\text{post},in}$) differ across sessions and bins because the simulated bias after recalibration $\beta_{i,\text{post}}$ depends on the adapter SOA fixed in session i and the simulation bin n.

Parameter estimation

We used the BADS toolbox (Acerbi & Ma, 2017) in MATLAB to optimize the set of parameters for the models because it outperforms fmincon when parameter numbers increase. We repeated each search 80 times with a different and random starting point to address the possibility of reporting a local minimum, and chose the parameter estimates with the maximum likelihood across the repeated searches.

References

Acerbi L., Ma W. J. (2017) **Practical bayesian optimization for model fitting with bayesian adaptive direct search** *Proceedings of the 31st International Conference on Neural Information Processing Systems* :1834–1844

Akaike H., Parzen E., Tanabe K., Kitagawa G. (1998) **Information theory and an extension of the maximum likelihood principle** *Selected papers of hirotugu akaike* :199–213

Aller M., Noppeney U. (2019) **To integrate or not to integrate: Temporal dynamics of hierarchical bayesian causal inference** *PLoS Biol* **17**

Badde S., Ley P., Rajendran S. S., Shareef I., Kekunnaya R., Röder B. (2020) **Sensory experience during early sensitive periods shapes cross-modal temporal biases** *Elife* **9**

Badde S., Navarro K. T., Landy M. S. (2020) **Modality-specific attention attenuates visual-tactile integration and recalibration effects by reducing prior expectations of a common source for vision and touch** *Cognition* **197**

Bedford F. L. (1999) **Keeping perception accurate** *Trends Cogn. Sci* **3**:4–11

Berniker M., Kording K. (2008) **Estimating the sources of motor errors for adaptation and generalization** *Nat. Neurosci* **11**:1454–1461

Burge J., Girshick A. R., Banks M. S. (2010) **Visual-haptic adaptation is determined by relative reliability** *J. Neurosci* **30**:7714–7721

Cao Y., Summerfield C., Park H., Giordano B. L., Kayser C. (2019) **Causal inference in the multisensory brain** *Neuron* **102**:1076–1087

Cary E., Lahdesmaki I., Badde S. (2024) **Audiovisual simultaneity windows reflect temporal sensory uncertainty** *Psychon. Bull. Rev*

Di Luca M., Machulla T.-K., Ernst M. O. (2009) **Recalibration of multisensory simultaneity: Cross-modal transfer coincides with a change in perceptual latency** *J. Vis* **9**:7–16

Dokka K., Park H., Jansen M., DeAngelis G. C., Angelaki D. E. (2019) **Causal inference accounts for heading perception in the presence of object motion** *Proc. Natl. Acad. Sci. U. S. A* **116**:9060–9065

Fain G. L. (2019) **Sensory transduction**

Fujisaki W., Shimojo S., Kashino M., Nishida S. (2004) **Recalibration of audiovisual simultaneity** *Nat. Neurosci* **7**:773–778

García-Pérez M. A., Alcalá-Quintana R. (2012) **On the discrepant results in synchrony judgment and temporal-order judgment tasks: A quantitative model** *Psychon. Bull. Rev* **19**:820–846

Grabot L., van Wassenhove V. (2017) **Time order as psychological bias** *Psychol. Sci* **28**:670–678

Hanson J. V. M., Heron J., Whitaker D. (2008) **Recalibration of perceived time across sensory modalities** *Exp. Brain Res* **185**:347–352

Harrar V., Harris L. R. (2008) **The effect of exposure to asynchronous audio, visual, and tactile stimulus combinations on the perception of simultaneity** *Exp. Brain Res* **186**:517–524

Heron J., Whitaker D., McGraw P. V., Horoshenkov K. V. (2007) **Adaptation minimizes distance-related audiovisual delays** *J. Vis* **7**:5–8

Hong F., Badde S., Landy M. S. (2021) **Causal inference regulates audiovisual spatial recalibration via its influence on audiovisual perception** *PLoS Comput. Biol* **17**

Keetels M., Vroomen J. (2007) **No effect of auditory-visual spatial disparity on temporal recalibration** *Exp. Brain Res* **182**:559–565

King A. J. (2005) **Multisensory integration: Strategies for synchronization** *Curr. Biol* **15**:R339–41

Körding K. P., Beierholm U., Ma W. J., Quartz S., Tenenbaum J. B., Shams L. (2007) **Causal inference in multisensory perception** *PLoS One* **2**

Kösem A., Gramfort A., van Wassenhove V. (2014) **Encoding of event timing in the phase of neural oscillations** *Neuroimage* **92**:274–284

Lennert T., Samiee S., Baillet S. (2021) **Coupled oscillations enable rapid temporal recalibration to audiovisual asynchrony** *Commun Biol* **4**

Locke S. M., Landy M. S. (2017) **Temporal causal inference with stochastic audio-visual sequences** *PLoS One* **12**

Machulla T.-K., Di Luca M., Froehlich E., Ernst M. O. (2012) **Multisensory simultaneity recalibration: Storage of the aftereffect in the absence of counterevidence** *Exp. Brain Res* **217**:89–97

McGovern D. P., Roudaia E., Newell F. N., Roach N. W. (2016) **Perceptual learning shapes multisensory causal inference via two distinct mechanisms** *Sci. Rep* **6**

Navarra J., Vatakis A., Zampini M., Soto-Faraco S., Humphreys W., Spence C. (2005) **Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration** *Brain Res. Cogn. Brain Res* **25**:499–507

O'Donohue M., Lacherez P., Yamamoto N. (2022) **Musical training refines audiovisual integration but does not influence temporal recalibration** *Sci. Rep* **12**

Pöppel E. (1988) **Mindworks: Time and conscious experience**

Roach N. W., Heron J., Whitaker D., McGraw P. V. (2011) **Asynchrony adaptation reveals neural population code for audio-visual timing** *Proc. Biol. Sci* **278**:1314–1322

Rohde M., Greiner L., Ernst M. O. (2014) **Asymmetries in visuomotor recalibration of time perception: Does causal binding distort the window of integration?** *Acta Psychol* **147**:127–135

Rohe T., Noppeney U. (2015) **Sensory reliability shapes perceptual inference via two mechanisms** *J. Vis* **15**

Roseboom W., Linares D., Nishida S. (2015) **Sensory adaptation for timing perception** *Proc. Biol. Sci* **282**

Samad M., Chung A. J., Shams L. (2015) **Perception of body ownership is driven by bayesian sensory inference** *PLoS One* **10**

Sato Y., Aihara K. (2011) **A bayesian model of sensory adaptation** *PLoS One* **6**

Sato Y., Toyoizumi T., Aihara K. (2007) **Bayesian inference explains perception of unity and ventriloquism aftereffect: Identification of common sources of audiovisual stimuli** *Neural Comput* **19**:3335–3355

Spence C., Squire S. (2003) **Multisensory integration: Maintaining the perception of synchrony** *Current Biology*

Stauffer C. C., Haldemann J., Troche S. J., Rammsayer T. H. (2012) **Auditory and visual temporal sensitivity: Evidence for a hierarchical structure of modality-specific and modality-independent levels of temporal information processing** *Psychol. Res* **76**:20–31

Tanaka A., Asakawa K., Imai H. (2011) **The change in perceptual synchrony between auditory and visual speech after exposure to asynchronous speech** *Neuroreport* **22**:684–688

Van der Burg E., Alais D., Cass J. (2013) **Rapid recalibration to audiovisual asynchrony** *J. Neurosci* **33**:14633–14637

Van der Burg E., Alais D., Cass J. (2015) **Audiovisual temporal recalibration occurs independently at two different time scales** *Sci. Rep* **5**

van Beers R. J., Wolpert D. M., Haggard P. (2002) **When feeling is more important than seeing in sensorimotor adaptation** *Curr. Biol* **12**:834–837

Vatakis A., Navarra J., Soto-Faraco S., Spence C. (2007) **Temporal recalibration during asynchronous audiovisual speech perception** *Exp. Brain Res* **181**:173–181

Vatakis A., Navarra J., Soto-Faraco S., Spence C. (2008) **Audiovisual temporal adaptation of speech: Temporal order versus simultaneity judgments** *Exp. Brain Res* **185**:521–529

Vroomen J., Keetels M. (2010) **Perception of intersensory synchrony: A tutorial review** *Atten. Percept. Psychophys* **72**:871–884

Vroomen J., Keetels M., de Gelder B., Bertelson P. (2004) **Recalibration of temporal order perception by exposure to audio-visual asynchrony** *Brain Res. Cogn. Brain Res* **22**:32–35

Wei K., Körding K. (2009) **Relevance of error: What drives motor adaptation?** *J. Neurophysiol* **101**:655–664

Wozny D. R., Beierholm U. R., Shams L. (2010) **Probability matching as a computational strategy used in perception** *PLoS Comput. Biol* **6**

Wozny D. R., Shams L. (2011) **Computational characterization of visually induced auditory spatial adaptation** *Front. Integr. Neurosci* **5**

Wozny D. R., Shams L. (2011) **Recalibration of auditory space following milliseconds of cross-modal discrepancy** *J. Neurosci* **31**:4607–4612

Yamamoto S., Miyazaki M., Iwano T., Kitazawa S. (2012) **Bayesian calibration of simultaneity in audiovisual temporal order judgments** *PLoS One* **7**

Yarrow K., Jahn N., Durant S., Arnold D. H. (2011) **Shifts of criteria or neural timing? the assumptions underlying timing perception studies** *Conscious. Cogn* **20**:1518–1531

Yarrow K., Minaei S., Arnold D. H. (2015) **A model-based comparison of three theories of audiovisual temporal recalibration** *Cogn. Psychol* **83**:54–76

Yarrow K., Solomon J. A., Arnold D. H., Roseboom W. (2023) **The best fitting of three contemporary observer models reveals how participants' strategy influences the window of subjective synchrony** *J. Exp. Psychol. Hum. Percept. Perform* **49**:1534–1563

Zaidel A., Turner A. H., Angelaki D. E. (2011) **Multisensory calibration is independent of cue reliability** *J. Neurosci* **31**:13949–13962

Editors

Reviewing Editor
Maria Chait
University College London, London, United Kingdom

Senior Editor
Barbara Shinn-Cunningham
Carnegie Mellon University, Pittsburgh, United States of America

Reviewer #1 (Public Review):

This study asks whether the phenomenon of crossmodal temporal recalibration, i.e. the adjustment of time perception by consistent temporal mismatches across the senses, can be explained by the concept of multisensory causal inference. In particular, they ask whether the explanation offered by causal inference better explains temporal recalibration better

than a model assuming that crossmodal stimuli are always integrated, regardless of how discrepant they are.

The study is motivated by previous work in the spatial domain, where it has been shown consistently across studies that the use of crossmodal spatial information is explained by the concept of multisensory causal inference. It is also motivated by the observation that the behavioral data showcasing temporal recalibration feature nonlinearities that, by their nature, cannot be explained by a fixed integration model (sometimes also called mandatory fusion).

To probe this the authors implemented a sophisticated experiment that probed temporal recalibration in several sessions. They then fit the data using the two classes of candidate models and rely on model criteria to provide evidence for their conclusion. The study is sophisticated, conceptually and technically state-of-the-art, and theoretically grounded. The data clearly support the authors' conclusions.

I find the conceptual advance somewhat limited. First, by design, the fixed integration model cannot explain data with a nonlinear dependency on multisensory discrepancy, as already explained in many studies on spatial multisensory perception. Hence, it is not surprising that the causal inference model better fits the data. Second, and again similar to studies on spatial paradigms, the causal inference model fails to predict the behavioral data for large discrepancies. The model predictions in Figure 5 show the (expected) vanishing recalibration for large delta, while the behavioral data don't' decay to zero. Either the range of tested SOAs is too small to show that both the model and data converge to the same vanishing effect at large SOAs, or the model's formula is not the best for explaining the data. Again, the studies using spatial paradigms have the same problem, but in my view, this poses the most interesting question here.

In my view there is nothing generally wrong with the study, it does extend the 'known' to another type of paradigm. However, it covers little new ground on the conceptual side.

On that note, the small sample size of n=10 is likely not an issue, but still, it is on the very low end for this type of study.

https://doi.org/10.7554/eLife.97765.1.sa2

Reviewer #2 (Public Review):

Summary:

Li et al.'s goal is to understand the mechanisms of audiovisual temporal recalibration. This is an interesting challenge that the brain readily solves in order to compensate for real-world latency differences in the time of arrival of audio/visual signals. To do this they perform a 3-phase recalibration experiment on 9 observers that involves a temporal order judgment (TOJ) pretest and posttest (in which observers are required to judge whether an auditory and visual stimulus were coincident, auditory leading or visual leading) and a conditioning phase in which participants are exposed to a sequence of AV stimuli with a particular temporal disparity. Participants are required to monitor both streams of information for infrequent oddballs, before being tested again in the TOJ, although this time there are 3 conditioning trials for every 1 TOJ trial. Like many previous studies, they demonstrate that conditioning stimuli shift the point of subjective simultaneity (pss) in the direction of the exposure sequence.

These shifts are modest - maxing out at around -50 ms for auditory leading sequences and slightly less than that for visual leading sequences. Similar effects are observed even for the longest offsets where it seems unlikely listeners would perceive the stimuli as synchronous

(and therefore under a causal inference model you might intuitively expect no recalibration, and indeed simulations in Figure 5 seem to predict exactly that which isn't what most of their human observers did). Overall I think their data contribute evidence that a causal inference step is likely included within the process of recalibration.

Strengths:

The manuscript performs comprehensive testing over 9 days and 100s of trials and accompanies this with mathematical models to explain the data. The paper is reasonably clearly written and the data appear to support the conclusions.

Weaknesses:

While I believe the data contribute evidence that a causal inference step is likely included within the process of recalibration, this to my mind is not a mechanism but might be seen more as a logical checkpoint to determine whether whatever underlying neuronal mechanism actually instantiates the recalibration should be triggered.

The authors' causal inference model strongly predicts that there should be no recalibration for stimuli at 0.7 ms offset, yet only 3/9 participants appear to show this effect. They note that a significant difference in their design and that of others is the inclusion of longer lags, which are unlikely to originate from the same source, but don't offer any explanation for this key difference between their data and the predictions of a causal inference model.

I'm also not completely convinced that the causal inference model isn't 'best' simply because it has sufficient free parameters to capture the noise in the data. The tested models do not (I think) have equivalent complexity - the causal inference model fits best, but has more parameters with which to fit the data. Moreover, while it fits 'best', is it a good model? Figure S6 is useful in this regard but is not completely clear - are the red dots the actual data or the causal inference prediction? This suggests that it does fit the data very well, but is this based on predicting held-out data, or is it just that by having more parameters it can better capture the noise? Similarly, S7 is a potentially useful figure but it's not clear what is data and what are model predictions (what are the differences between each row for each participant; are they two different models or pre-test post-test or data and model prediction?!).

I'm not an expert on the implementation of such models but my reading of the supplemental methods is that the model is fit using all the data rather than fit and tested on held-out data. This seems problematic.

I would have liked to have seen more individual participant data (which is currently in the supplemental materials, albeit in a not very clear manner as discussed above).

The way that S3 is described in the text (line 141) makes it sound like everyone was in the same direction, however, it is clear that 2 /9 listeners show the opposite pattern, and 2 have confidence intervals close to zero (albeit on the -ve side).

https://doi.org/10.7554/eLife.97765.1.sa1

Reviewer #3 (Public Review):

Summary:

Li et al. describe an audiovisual temporal recalibration experiment in which participants perform baseline sessions of ternary order judgments about audiovisual stimulus pairs with various stimulus-onset asynchronies (SOAs). These are followed by adaptation at several adapting SOAs (each on a different day), followed by post-adaptation sessions to assess changes in psychometric functions. The key novelty is the formal specification and

application/fit of a causal-inference model for the perception of relative timing, providing simulated predictions for the complete set of psychometric functions both pre and post-adaptation.

Strengths:

(1) Formal models are preferable to vague theoretical statements about a process, and prior to this work, certain accounts of temporal recalibration (specifically those that do not rely on a population code) had only qualitative theoretical statements to explain how/why the magnitude of recalibration changes non-linearly with the stimulus-onset asynchrony of the adaptor.

(2) The experiment is appropriate, the methods are well described, and the average model prediction is a fairly good match to the average data (Figure 4). Conclusions may be overstated slightly, but seem to be essentially supported by the data and modelling.

(3) The work should be impactful. There seems a good chance that this will become the go-to modelling framework for those exploring non-population-code accounts of temporal recalibration (or comparing them with population-code accounts).

(4) A key issue for the generality of the model, specifically in terms of recalibration asymmetries reported by other authors that are inconsistent with those reported here, is properly acknowledged in the discussion.

Weaknesses:

(1) The evidence for the model comes in two forms. First, two trends in the data (non-linearity and asymmetry) are illustrated, and the model is shown to be capable of delivering patterns like these. Second, the model is compared, via AIC, to three other models. However, the main comparison models are clearly not going to fit the data very well, so the fact that the new model fits better does not seem all that compelling. I would suggest that the authors consider a comparison with the atheoretical model they use to first illustrate the data (in Figure 2). This model fits all sessions but with complete freedom to move the bias around (whereas the new model constrains the way bias changes via a principled account). The atheoretical model will obviously fit better, but will have many more free parameters, so a comparison via AIC/BIC or similar should be informative.

(2) It does not appear that some key comparisons have been subjected to appropriate inferential statistical tests. Specifically, lines 196-207 - presumably this is the mean (and SD or SE) change in AIC between models across the group of 9 observers. So are these differences actually significant, for example via t-test?

(3) The manuscript tends to gloss over the population-code account of temporal recalibration, which can already provide a quantitative account of how the magnitude of recalibration varies with adaptor SOA. This could be better acknowledged, and the features a population code may struggle with (asymmetry?) are considered.

(4) The engagement with relevant past literature seems a little thin. Firstly, papers that have applied causal inference modelling to judgments of relative timing are overlooked (see references below). There should be greater clarity regarding how the modelling here builds on or differs from these previous papers (most obviously in terms of additionally modelling the recalibration process, but other details may vary too). Secondly, there is no discussion of previous findings like that in Fujisaki et al.'s seminal work on recalibration, where the spatial overlap of the audio and visual events didn't seem to matter (although admittedly this was an N = 2 control experiment). This kind of finding would seem relevant to a causal inference account.

References:
Magnotti JF, Ma WJ and Beauchamp MS (2013) Causal inference of asynchronous audiovisual speech. Front. Psychol. 4:798. doi: 10.3389/fpsyg.2013.00798
Sato, Y. (2021). Comparing Bayesian models for simultaneity judgement with different causal assumptions. J. Math. Psychol., 102, 102521.

(5) As a minor point, the model relies on simulation, which may limit its take-up/application by others in the field.

(6) There is little in the way of reassurance regarding the model's identifiability and recoverability. The authors might for example consider some parameter recovery simulations or similar.

(7) I don't recall any statements about open science and the availability of code and data.

https://doi.org/10.7554/eLife.97765.1.sa0