The organization of experiences and its effects on episodic memory and decision-making

Yeon Soon Shin

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Princeton Neuroscience Institute
Advisers: Yael Niv and Kenneth A. Norman

September 2020

Abstract

Idiosyncratic experiences from the past guide our future predictions and decisions in meaningful ways. We generalize what we learned from the past to similar situations. This dissertation investigates how we group our experiences in a way that supports effective retrieval of relevant experiences and generalization from those experiences. I hypothesize that hidden causal structures that are assumed to have generated a set of observable events, called latent causes, serve as a meaningful basis for generalization. In Chapter 2, I first test whether external environmental contexts that have unique sets of features generated from distinct latent causes organize episodic memories. In this study, subjects learned a list of words while performing context-appropriate tasks and recalled the words either in the same or in a different context. I demonstrate that recalling memories in the original learning environmental context facilitates episodic memory retrieval. This effect is larger when the words were judged to be conceptually more relevant to the context, suggesting that integration to the latent causes is important in organizing memories of experiences. To further this idea, in Chapter 3, I propose a framework for grouping and segmenting events in episodic memory that is based on inferring latent causes. This framework can reconcile seemingly contradictory empirical findings, such as memory biases towards both extreme episodes and the average of episodes, by sampling memories within a latent cause (item-level sampling) or across latent causes (cluster-level sampling). In Chapter 4, I test the predictions of cluster-level sampling in social decision making, in which impressions about a group are biased as experiences with the group members are summarized at the level of inferred latent causes. In the Conclusion, I discuss an ongoing project that aims to directly measure the latent-cause inference process and map cognitive constructs onto model parameters. I propose how this latent-cause inference can provide a framework for social cognition where key features such as current beliefs, mental states, and traits are hidden.

Acknowledgements

I would like to extend my deepest gratitude to my advisers, Yael Niv and Ken Norman, for their unwavering support over the course of PhD. This dissertation certainly would not have been possible without the guidance they graciously offered from the inception of the projects. Their sharpest scientific intuition, which continues to amaze me even after six years of working with them, has made even the most complicated models and concepts more accessible. I feel deeply grateful to have had the opportunity to grow up as a scientist with two most incredible scientists as my advisers. I hope to pay it forward in some form or another, although I know that it would be a challenge to reach their level of dedication to mentorship.

I would also like to thank my collaborators/friends/lab mates/and beyond. I honestly do not think that a single category label is enough for most of them, which makes this a really difficult clustering problem. If you are one of those people, I believe you will know that you are a member of multiple clusters for me, and that everything I say below applies to you, even though you are arbitrarily categorized as one.

I am immensely thankful for my brilliant collaborators, Sarah DuBrow, Rolando Masís-Obando, Neggin Keshavarzian, Riya Davé, Dan-Mircea Mirea, and Paul Kazelis, who made the projects infinitely more fun and stimulating, whether we are in a VR room in the PNI basement or talking over hangout/zoom. I am also greatly indebted to my friends at PNI, Nina Rouhani, Angela Radulescu, Lili Cai, Alex Libby, Aaron Kurosu, and many, many others. Even if we did not directly work on projects together, talking science with them has been invaluable and inspiring. I am enourmously thankful for my support network, Jeayoon Lee, Alice Yoon, DongWon Oh, Susie Kim, and especially Liza Dzul. They have given me strength to become the person I am today. Finally, I would like to thank my family for all their love and support. I dedicate this dissertation to them.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

There are no two events that can be exactly the same, as at least one aspect–time–is constantly moving forward. Nonetheless, these idiosyncratic experiences from the past guide our future predictions and decisions in meaningful ways. We generalize what we learned from the past to similar situations. For instance, when meeting a new person, we may draw from previous encounters with other people who share similar qualities with the person. In drawing from the past, which specific encounters come to mind among the myriad of experiences we have accumulated throughout life? How do we group our experiences in a way that supports effective retrieval of relevant experiences and generalization from those experiences?

Context, defined as characteristics of the external or mental environment (Eich et al., 1975; Underwood, 1977), can be the basis for the organization of the experiences. In a classic study that investigated the effect of external environment in episodic memory, Godden and Baddeley (1975) asked scuba divers to study a list of words underwater or on land, and tested their memory either in the same or a different context. The divers who recalled the words in the original study environment recalled more words than the ones who were tested in a different context, suggesting that being in the same context facilitates the retrieval of memories that belong to the context.

Context also influences learning in animals. For example, a rodent who has learned that a tone predicts a shock in a cage shows a fear response upon their return to the same cage, even after going through extinction in a different cage (Bouton and Bolles, 1979; Gershman et al., 2010). In experimental settings, context is often defined as external physical environments as in the aforementioned examples, as the environmental contexts are what the experimenters can easily manipulate. However, physical surroundings are not always meaningful in guiding what subjects find relevant in a given situation. Studies that only manipulate the environmental context have mixed results (for a review, see Smith and Vela, 2001). Indeed, even in the same physical location, we can find ourselves in very different situations. For example, if we are asked to give a talk at our alma mater's lecture hall, it would be more helpful to remember what happened in a similar situation (e.g., giving a talk at a different institution) than events that occurred in the same lecture hall (e.g., taking a class). This idea is in line with studies that show mentally reinstating the learning context, without revisiting the physical site, is enough to bring back the memories from the context (Smith, 1979). Even when external cues do not change, our internal state may change, providing a different context for processing incoming information and retrieving old information. What is it, then, that enables efficient retrieval and generalization of our experiences?

"Latent causes" – hidden causal structures that are assumed to generate a set of observable events, serve as a meaningful basis for generalization (Courville et al., 2005; Gershman et al., 2010). For instance, when an event shares its latent cause (e.g., giving a talk) with previous events, retrieving those events and generalizing from them would be normative. In this thesis, I investigate how the hidden structure of causes organizes episodic memories, segments a continuous stream of experiences, and summarizes experiences to support decision-making.

In **Chapter 2**, I demonstrate that recalling memories in the original learning context facilitates episodic memory retrieval. As we cannot directly manipulate what latent cause is inferred by subjects, we manipulate a set of features that would encourage subjects to form and activate distinct latent causes. Specifically, we use two semantically-distinct contexts (underwater and Mars), each associated with a separate set of actions in a virtual reality task. We show that items are better recalled later when they are retrieved in the same context where they were learned. We also show that this memory benefit is larger for items deemed context-relevant at encoding, potentially due to the importance of integrating items into an active latent cause for the successful generation of this effect.

In **Chapter 3**, I propose that a continuous stream of experiences can be segmented and clustered by inferred latent causes. I review event segmentation theory, overviewing how event segmentation influences ongoing processing, subsequent memory retrieval, and decision making, as well as some proposed underlying mechanisms. I then explore how inferences, or beliefs, about what generates our experience (i.e., latent cause) may be the foundation of event cognition. In this inference-based framework, experiences are grouped together according to their latent causes. Segmentation then occurs when the inferred latent causes change, creating an event boundary. This framework offers an alternative to dominant theories of event segmentation, allowing boundaries to occur independent of perceptual change and even when transitions between latent causes are predictable. I describe how this framework can reconcile seemingly contradictory empirical findings (for example, the fact that memory can be biased towards both extreme episodes and the average of episodes).

In **Chapter 4**, I investigate how we use latent causes to evaluate other people. In particular, I propose that when we evaluate a group after observing its members' behaviors, biases can emerge when we infer the latent causes of the observed behaviors. I use a Bayesian latent-cause inference model that learns environmental statistics, combining highly similar events together and separating rare or highly variable observations. The model predicts that group evaluations that rely on averaging inferred latent causes will overweight variable events. I empirically test these model-derived predictions in eight decision-making experiments, in which subjects observe a sequence of social or non-social behaviors and are subsequently asked to estimate the average of observed values. As predicted by a latent-cause model, when we queried for the average estimate after subjects saw all events, the estimate was biased toward rare and highly variable events. When subjects tracked a summary value as observations of events accrued, instead of parsing events into distinct latent causes, the bias was eliminated. These results suggest that biases in evaluations of social groups, such as the negativity bias, may arise from inferring the hidden causes of group members' behaviors.

In the **Conclusion**, I discuss an ongoing project that aims to measure the latent-cause inference process directly and map cognitive constructs onto model parameters. I propose how this latent-cause inference can provide a framework for social cognition where key features such as current beliefs, mental states, and traits are hidden.

Chapter 2

Retrieval of episodic memory and latent cause

The contents of this chapter were published in Shin, Y.*, Masís-Obando, R.*, Keshavarzian, N., Davé, R., & Norman, K. (2020, April 6). Context-dependent memory effects in VR. https://doi.org/10.31234/osf.io/prdb7.

2.1 Introduction

Returning to an alma mater for a reunion can bring back memories from the past. Walking by campus may make it easier to recall past events that took place in the dorms, classrooms, and dining halls, even when those memories are not easily retrievable elsewhere. This flood of memories when returning to an old environment is known as the environmental reinstatement effect (Smith, 1979). Research in episodic memory explains this effect with the encoding-specificity hypothesis, which posits that increasing levels of overlap with the encoding context during retrieval aids memory performance (Tulving and Thomson, 1973). However, the beneficial effect of context reinstatement in recall has not been consistently supported in the memory literature (for a review, see Smith and Vela, 2001). Contexts have been manipulated in various ways such as background colors (Weiss and Margolius, 1954; Isarida and Isarida, 2007) and physical rooms (Eich, 1985; Fernandez and Glenberg, 1985), but these manipulations do not always lead to a context-reinstatement effect (Isarida and Isarida, 2007; Fernandez and Glenberg, 1985; Wälti et al., 2019). The discrepancy between the strong anecdotal psychological experience and weak experimental evidence indicates that extant experimental paradigms are missing key features that are responsible for evoking context-dependent memory in real life. What, then, are these missing features?

One of the seminal studies that showed strong context-dependency used rich, real-world environments to manipulate the congruency between encoding and retrieval contexts (Godden and Baddeley, 1975). They found that scuba divers recalled learned words better when the retrieval context (underwater or land) matched the encoding context. It is worth noting, however, that subjects were put into vastly different situations that likely activated different bodies of knowledge associated with each environment (e.g., how to swim and breathe underwater). Similarly, Smith and Manzano (2010) showed a robust context reinstatement effect in a study where contexts were manipulated by video scenes showing situations that subjects were likely to be already familiar with. Relatedly, a change in the mental representation of the current situation can reduce access to episodes that happened before the change (for a review, see DuBrow et al., 2017). These studies suggest that, in order to demonstrate a robust context effect, subjects should mentally represent distinct situations and activate distinct sets of knowledge while encoding and retrieving the items. Standard laboratory experiments that merely change the physical environment or the color on a screen may have failed to elicit the effect because they failed to make subjects believe that they were in different situations.

These ideas fit with prior work in cognitive psychology (Bransford and Johnson, 1972; Alba and Hasher, 1983) and cognitive neuroscience (Poppenk and Norman, 2012; Whittington et al., 2019; Tse et al., 2007; Gilboa and Marlatte, 2017) showing that activation of relevant pre-existing knowledge (i.e., schemas) can facilitate new learning by providing a "scaffold" onto which new information can be attached. Once information has been attached to this contextual scaffold, reinstating the scaffold at test should facilitate recall, and taking it away should hurt recall, thereby leading to a context change effect.

In the present study, we aimed to develop an experimental paradigm that can produce a strong context reinstatement effect in a laboratory setting. First, we used two virtual reality (VR) environments that we predicted would activate distinctive pre-existing sets of knowledge (i.e. schemas) in the subjects, underwater (UW) and Mars planet (MP) environments. In other words, these two environments differed not only perceptually, but also in their prior conceptual associations. Furthermore, to maximize the difference between the two contexts while expanding on their existing knowledge, we also had subjects perform context-specific actions that were physically distinctive: Subjects performed downward motions in UW and upward motions in MP when interacting with objects to initiate each session (i.e., context-initiation action sequences) and to discover to-be-remembered word items (i.e., item-finding actions; Fig. 2.1b). Second, in order for subjects to have these bodies of knowledge readily available at the time of study and test, as was the case for the divers in Godden and Baddeley (1975), we familiarized subjects with the virtual environments in a foraging task where they collected objects dispersed throughout each environment. This was followed by a practice of the associated sequences of actions (context-initiation action sequences and item-finding actions) before they went into the main task (Fig. 2.1c top row).

Another key desideratum is that, during encoding, these distinct bodies of knowledge should be activated when performing the experimental tasks. If subjects are aware of a future memory test, they may use mnemonic strategies unrelated to the present context and ignore other inputs from the environment. Thus, to ensure contextual relevance during encoding and to conceal the purpose of the study, we used a surprise memory test. Furthermore, a cover task for encoding forced subjects to deliberately integrate the memory items into the given context (Eich, 1985). Specifically, we asked subjects to judge whether a shown item was useful in the context where it was found. We hoped that these judgment tasks that were unique to each environment, paired with the item-finding and context-initiation action sequences (also unique to each environment), would foster the activation of context-unique associations during encoding. Additionally, we hypothesized that items that were affirmatively judged to be useful in the encoding environment would show a stronger context-change effect – the idea being that schema-congruent items would be more strongly integrated into the active schema at encoding and thus suffer more if that schema were not active at retrieval (Bransford and Johnson, 1972). The context-initiation action sequences were also performed at the beginning of the retrieval session, making the encoding and retrieval sessions more similar for the "same" condition and more distinctive for the "different" condition.

Lastly, we also manipulated the interval between encoding and retrieval. When items were encountered very recently, they may be retrievable without relying on contextual cues, and this may weaken context dependency (Smith and Vela, 2001). To test this, we used two levels of encoding-retrieval intervals, where a longer interval (the “delay” condition) was expected to produce a stronger reinstatement effect than a shorter interval (the “immediate” condition).

In summary, our approach was to combine as many factors as we could in the service of ensuring that participants activated distinct bodies of knowledge (i.e., schemas) when learning word lists in the two contexts. The benefit of this combined approach is to maximize effect size (if our hypothesis is correct), with the complementary drawback that – if we obtain an effect – we are not in a position to say which of the factors are necessary and sufficient for driving the effect.

2.2 Methods

2.2.1 Participants

Seventy-two adults (50 female; 22 male) recruited from Princeton University and the university community participated in our study. All but two participants were right-handed (1 left-handed, 1-ambidextrous). Informed consent was obtained from all participants in accordance with Princeton Institutional Review Board, and subjects were each provided with monetary compensation or course credit for participating in the study. With the exception of 8 participants who did not complete the study due to technical issues with the VR devices and/or Unity (e.g. VIVE wireless disconnected or screen froze during encoding), a total of 64 subjects were included in the analyses.

2.2.2 Task and Procedure

We used a 2 (study-test context congruency: same vs. different) x 2 (study-test interval: immediate vs. delay) between-subjects design (Fig. 2.1a). The experimental procedure was divided into three parts: a training phase with sessions in both UW and MP (Fig. 2.1c top row), two encoding sessions in one of the two environments (Fig. 2.1c middle row), and a retrieval session either in the original encoding environment (the "same" condition) or in the other environment (the "different" condition; Fig. 2.1c bottom row). In the "immediate" condition, encoding and retrieval occurred on the same day, while for the "delay" condition, retrieval took place the day after encoding.

Training

The training session is depicted in Fig. 2.1c, top row. After signing consent and screening forms, subjects were handed a paper with a fictionalized mission statement outlining the task to be performed in VR. Subjects were told that they were pioneers developing alternative places for humans to inhabit in the future and instructed to judge whether they should keep the items they discover based on the usefulness and pertinence to thriving in the environment they were in. This was meant to obscure the purpose of the study and the surprise memory test later.

After reading the mission statement, subjects were familiarized with the VR software by entering a practice environment. In this environment, subjects learned how to navigate and interact with objects that they would encounter in the main task. Subjects learned how to navigate by teleportation and how to find and judge items for the cover task (e.g., bending down, reaching upwards, etc.). Experimenters communicated with subjects during training via an intercom system connected to the head-mounted display audio-device.

Subjects were then introduced to the MP and UW environments. The order of the environments was counterbalanced across subjects. In each environment, subjects first performed a foraging task where subjects explored and collected 20 floating spheres scattered across the environment. This was followed by a context-initiation action sequence (Fig. 2.1b top row), which subjects would be required to perform at the beginning of each session. For this, subjects in UW needed to bend down to reach a key and insert it into the key-hole of a large treasure chest in the center of the environment, while subjects in MP needed to reach upwards to grab a floppy-disk and insert it to the side of a podium near the center of the crater. After the context-initiation action sequence, subjects heard an audio instruction for the task they needed to perform in the session. The instruction guided subjects to practice item-finding actions and the judgment mechanics that required reaching out and picking either a red or green cube with their controller. The item-finding actions were also unique to the environment. Subjects in UW found a to-be-remembered item by bending down and digging inside of a chest, whereas MP subjects did so by reaching up to a floating rock and scanning it while it was latched onto the scanner (Fig. 2.1b middle row). After finding and judging four items in the environment, they were teleported out and asked to perform two types of distractor tasks. The first distractor task was a countdown task in which subjects were instructed to count down from 100 to negative 300 by a randomly-generated single-digit number in an empty environment. The second distractor task was a monster-smash task that required subjects to hit as many monster heads off the ground as possible. The countdown task and monster-smash task both lasted 1 minute and were always performed in succession via virtual transportation from the countdown room to the monster smashing platform. Subjects repeated these tasks when trained in the second environment.

Encoding

After the training phase was completed, subjects were either transported to MP or UW for the encoding sessions. There were two encoding sessions (Fig. 2.1c middle row). At the beginning of the encoding session, subjects first performed the context-initiation task that was specific to the environment, after which they listened to the recorded instruction. Subjects then performed a cover task where they made judgments about whether the items they discovered should be kept in the environment based on their perceived usefulness for surviving in that environment. There were 24 items in each session, and they were told that they should only keep roughly half of the items. Immediately after the word item disappeared, the judgment cubes (green for useful, red for harmful) appeared in front of the subjects' virtual visual field simultaneously and remained there for 6 seconds before disappearing. If no selection was made within the 6 s trial window, the judgment trial was counted as missed.

After successfully discovering all 24 items in chests (in UW) or rocks (in MP), subjects performed distractor tasks (i.e., countdown and monster-smash tasks). Once both distractors were completed, subjects were virtually transported back to the same encoding environment as the first encoding session and initiated the second encoding session. The second encoding session was identical to the first except for the to-be-remembered items (24 new words). After the second encoding session, subjects again performed the two distractor tasks.

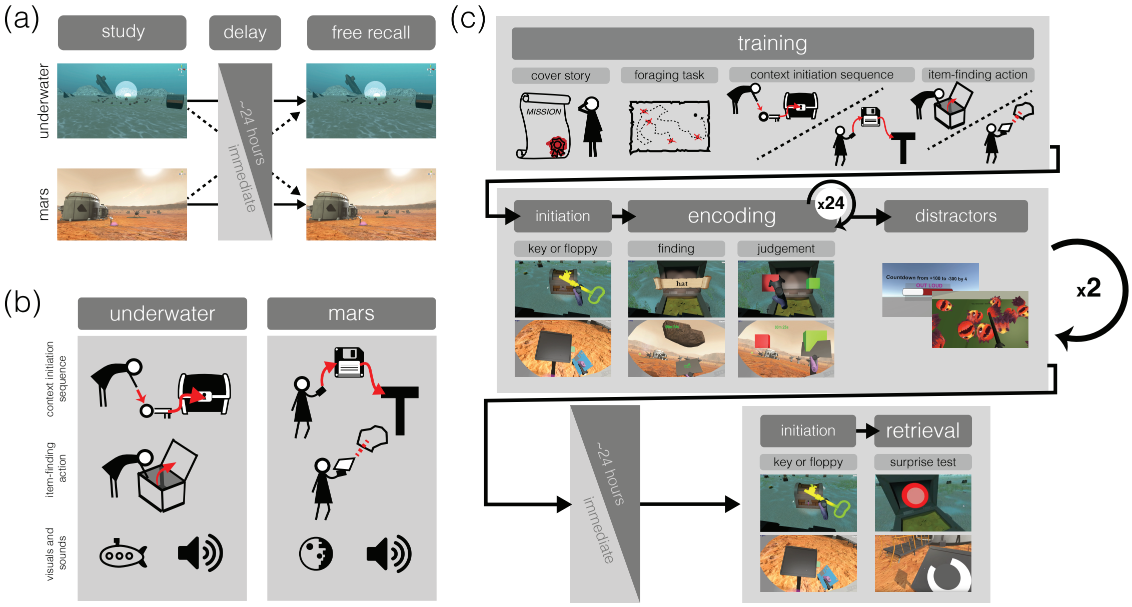

Figure 2.1: Methods pipeline (a) Study design. We used a 2 (study-test context congruency: same vs. different) x 2 (study-test interval: immediate vs. delay) between-subjects design. Subjects were split between getting tested in either the same environment where learning occurred or in the other environment. Dotted arrows indicate the "different" condition while solid arrows indicate the "same" condition. Subjects went into the free recall session either immediately or approximately 24 hours after learning of the items occurred. **(b) Environment-specific gameplay mechanics and stimuli.** Underwater (UW) and Mars planet (MP) each contained context-specific gameplay mechanics (top and middle rows), with distinct visual and auditory stimuli (bottom row). **(c) Task procedure** Our paradigm consisted of three main phases (delineated by the three rows). Subjects first read the cover story and were familiarized with each of the environments (top row) through the following tasks: (1) a foraging task, (2) the context-specific initiation sequence, and (3) item-finding actions. Each practice run was followed by the two distractor tasks. In the second phase (middle row), each subject went through two encoding sessions in one of the two environments. Before encoding, subjects performed the context-initiation sequence specific to the environment. They then found to-be-remembered items by performing the environment-unique item-finding action and judged whether the item was useful or not for survival in the present environment. They did this for 24 items in a given session before proceeding to two distractor tasks (right most middle row). The encoding session was performed twice in the same environment with two non-overlapping sets of word items. Either immediately following the encoding session or after a one-day delay (bottom row), subjects again performed the context-initiation sequence specific to the retrieval environment, at which point they were asked to recall all the words they had judged during study (i.e., 48 words).

Retrieval

Following the encoding task, subjects were introduced to the retrieval session (Fig. 2.1c bottom row). Subjects in the “immediate” condition continued to the retrieval task while subjects in the “delay” condition were told to return the next day for additional missions. Before the retrieval session, subjects in the “delay” condition were re-familiarized with the retrieval environment and performed the two distractor tasks. Subjects in the “immediate” condition were transported to the retrieval environment immediately after the distractor tasks that followed the second encoding session.

The “same” condition subjects returned to the same environment in which encoding took place for retrieval (i.e., UW-UW or MP-MP), while the “different” condition subjects faced the surprise memory test in the environment that differed from the encoding session (i.e., UW-MP or MP-UW).

Once subjects were virtually transported to the retrieval environment, subjects performed the corresponding context-initiation task after which they learned for the first time that there was a surprise memory test. Subjects were instructed to verbally recall all the word items they had discovered from both encoding sessions (i.e., 48 words), regardless of whether items were judged to be useful or not. The recall period lasted 2 minutes.

2.2.3 Materials

Environments

Two distinctive VR environments were custom-built and explored by subjects with a wireless head-mounted virtual reality display (HMD): underwater (UW) and Mars planet (MP). To maximize the difference between the two contexts, each had a different layout and a set of thematically corresponding specific player actions, objects, and sounds (Fig. 2.1b). Sound effects were obtained online or made in-house with Ableton Live software and instruction audios were created using a text-to-speech service (fromtexttospeech.com).

The UW environment resembled the ocean floor. The landscape was tinted blue with coral surrounding the play area, and 3-D models of shipwrecks, submarines, boats, and chests were scattered around the environment. Items to be found were words located inside small chests; to interact with the chests, subjects had to bend down, lift the lid, and place their controller inside the chest to "dig" into the chest to reveal the word to judge. In addition to the bubble and underwater sounds, all instructions in the UW context were given by a male voice.

The MP environment resembled science fiction depictions of the planet Mars. The landscape was tinted orange with mountains far in the distance of a large crater which served as the play area. Models of spaceships, satellites, and floating rocks were interspersed around the crater with additional spaceships hovering in the sky. In MP, to interact with the floating rocks, subjects had to lift the rocks with their controller and press the trigger button to "scan" them and reveal the word. To maximize immersion, coarse wind sounds played in the background, and static sound effects were triggered with every word discovery (e.g. "scanning rock"). All instructions in the MP context were given by a female voice.

Words

Forty-eight concrete noun words were used as encoding items. The same set of 48 words was presented for both MP and UW. To ensure that roughly half of the items were judged useful in both environments, we normed the words using Amazon Mechanical Turk, where the context was given by a screenshot of the corresponding VR environment and 180 words were judged for usefulness in the given environment. The mean probability of the chosen set of words being judged useful was 0.56 ($SD = 0.28$) in MP and 0.48 ($SD = 0.25$) in UW.

2.2.4 Apparatus

All tasks were presented on a wireless HTC Vive Pro head-mounted display (1440 × 1600 resolution per eye, with a 90 Hz refresh rate, and built-in headphones and integrated microphone) which was wirelessly connected (with a HTC Wireless Adapter) to a custom-built computer running 64-bit Windows 10 on an Intel Core i7-7800X CPU @ 3.50GHz with 32GB RAM and an Nvidia Geforce RTX 2080 graphics card.

All tasks were created and coded in Unity3D 2017.4.3, a game-development platform, with Virtual Reality Toolkit (VRTK; vrtk.com), a virtual-reality programming tool-kit for Unity3D. 3D models, textures, environments and other assets were downloaded from the Unity Asset Store (assetstore.unity.com), Turbosquid (turbosquid.com) and then modified or custom-built using Blender (blender.org).

2.2.5 Statistical Analyses

We performed statistical analyses using a generalized linear mixed-effects model in R with the “lme4” package (Bates et al., 2015), treating subjects and word stimuli as random effects. For confidence intervals, we performed bootstrapping, sampling different sets of subjects with replacement 5000 times.

2.3 Results

2.3.1 Encoding

Overall, subjects spent 7.44 minutes per encoding session (SD = 0.92 min, 95% bootstrap CI $[7.23, 7.68]$). Subjects spent more time in MP (M = 7.89 min, SD = 0.99 min, 95% bootstrap CI $[7.57, 8.25]$) than in UW (M = 7.00 min, SD = 0.58 min, 95% bootstrap CI $[6.80, 7.20]$; $t(49.83) = 4.380$, $p < 0.001$), reflecting longer distance (in arbitrary units) traveled in MP (M = 6.13, SD = 0.49, 95% bootstrap

CI $[5.97, 6.31]$) than in UW ($M = 5.00$, $SD = 0.55$, 95% bootstrap CI $[4.81, 5.19]$); $t(61.29) = 8.714$, $p < 0.001$).

For the decision task, 52% ($SD = 9\%$, 95% CI $[50\%, 55\%]$) of the items were judged as useful for the environment. Mixed-effects logistic regression analysis showed that there was no significant difference in the probability of judging items useful between encoding environments ($\beta = 0.098$, $SE = 0.107$, Wald $Z = 0.912$, $p = 0.362$; $M = 0.51$, $SD = 0.10$, 95% bootstrap CI $[0.48, 0.55]$; UW $M = 0.54$, $SD = 0.08$, 95% bootstrap CI $[0.51, 0.56]$). Decision reaction times did not significantly differ between "Useful" ($M = 793.25$ ms, $SD = 271.27$ ms, 95% bootstrap CI $[728.18, 860.91]$) and "Harmful" ($M = 868.58$ ms, $SD = 267.64$ ms, 95% bootstrap CI $[805.36, 934.33]$; $t(125.98) = -1.582$, $p = 0.116$) judgments, or between encoding environments ($t(61.25) = 1.510$, $p = 0.136$; MP $M = 864.91$ ms, $SD = 223.25$ ms, 95% bootstrap CI $[787.01, 943.16]$; UW $M = 775.54$ ms, $SD = 249.55$ ms, 95% bootstrap CI $[689.70, 864.36]$).

2.3.2 Retrieval

Overall accuracy was 0.29 ($SD = 0.12$, 95% bootstrap CI $[0.26, 0.32]$). First, we tested the context reinstatement effect, the benefit in retrieving memory items in the same environment as the encoding environment as opposed to a different environment. Given that a mixed-effects logistic regression model that controlled for stimulus effect (i.e., word items) as a random effect showed a better fit ($BIC = 3416.433$) than a model that did not ($BIC = 3646.219$), we used a model with subjects and word items as random effects to predict whether an item is recalled as a function of the context congruency manipulation. The mixed-effects model showed a significant benefit for the "same" condition ($M = 0.32$, $SD = 0.11$, 95% bootstrap CI $[0.28, 0.36]$) compared to when recall was in a different environment ($M = 0.26$, $SD = 0.11$, 95% bootstrap CI $[0.22, 0.30]$); $\beta = 0.384$, $SE = 0.160$, Wald $Z = 2.406$, $p < 0.05$; Fig. 2.2a), suggesting that retrieving items is easier when the retrieval environment matches the encoding environment.

To investigate additional factors that affect context reinstatement effect, we tested whether the effect depended on the interval between encoding and retrieval. Although the interaction between context congruency and interval was not significant ($\beta = 0.149$, $SE = 0.251$, Wald $Z = 0.595$, $p = 0.552$) when tested using a mixed-effects logistic regression model, the effect size for delay condition (Cohen's D = 0.917) was larger than the immediate condition (Cohen's D = 0.599; Fig. 2.2b). There was a main effect of interval where delayed recall performance was significantly worse than immediate recall ($\beta = -0.781$, $SE = 0.126$, Wald $Z = -6.205$, $p < 0.001$).

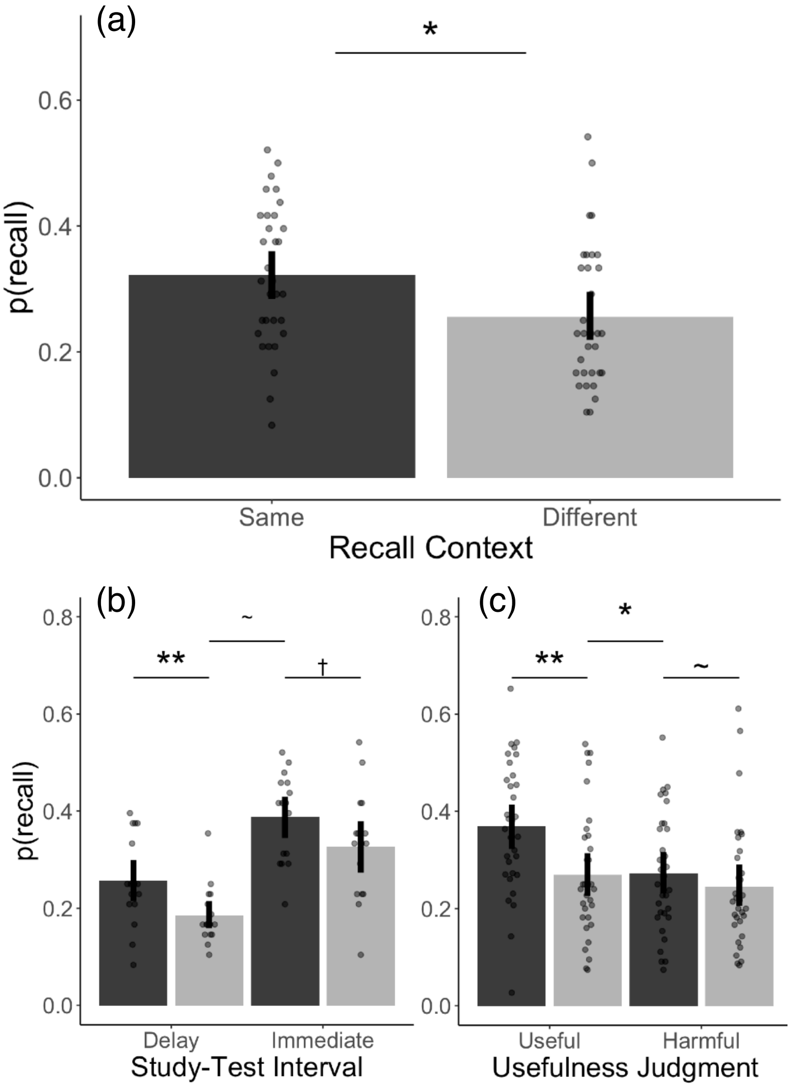

Figure 2.2: Recall performance (a) The context reinstatement effect. Subjects recalled significantly more words when the recall context was the same as the encoding context (dark gray; N = 32) than when it was different (light gray; N = 32). **(b) Recall performance as a function of context congruency and study-test interval.** The interaction between context congruency and study-test interval was not significant. In the "delay" condition, there was a significant benefit in recalling items in the same context (dark gray; N = 16) compared to the different context (light gray; N = 16). This benefit was not significant in the "immediate" condition (the "same" condition N = 16; the "different" condition N = 16) **(c) Recall performance as a function of context congruency and usefulness judgment.** There was a significant interaction between context congruency and usefulness judgment, where useful items showed a greater benefit from context congruency. Note: Dots indicate individual subjects. Error bars indicate 95% bootstrap confidence intervals. $**p < 0.01$, $*p < 0.05$, $\dagger p < 0.1$, $\sim n.s.$

To further explore the mechanisms by which context reinstatement affects memory recall, we looked at the relationship between the usefulness judgment and context congruency. A mixed-effects logistic regression model showed an interaction between context congruency and usefulness judgment ($\beta = 0.375$, $SE = 0.178$, Wald $Z = 2.106$, $p < 0.05$), as well as main effects of context congruency ($\beta = 0.344$, $SE = 0.163$, Wald $Z = 2.108$, $p < 0.05$) and usefulness judgment ($\beta = 0.371$, $SE = 0.100$, Wald $Z = 3.714$, $p < 0.001$; Fig. 2.2c). Planned comparisons showed that the context reinstatement effect was significant among the "useful" items ($\beta = 0.521$, $SE = 0.180$, Wald $Z = 2.898$, $p < 0.01$), but was not significant among the "harmful" items ($\beta = 0.177$, $SE = 0.193$, Wald $Z = 0.918$, $p = 0.359$). These results suggest that reinstating the context preferentially brings back memories for the contextually relevant items, thereby boosting overall memory performance.

2.4 Discussion

In this study, we investigated the environmental context-dependent memory effect in virtual reality, with a design intended to activate distinct bodies of context-specific knowledge. Subjects studied items either underwater (UW) or on Mars planet (MP), under a cover story in which they judged usefulness of items for the given context, and took a surprise free recall test in either the same or different context as the study context. We showed that the items were better recalled when retrieved in the same context as the study context. Importantly, these context-dependent memory effects were only obtained for items that were judged to be useful for survival in the encoding environment.

Circling back to our central motivation, why has it been so difficult to replicate context-dependent memory effects in the laboratory? In the introduction, we hypothesized that the key to obtaining context-dependent memory effects was to integrate items into different schemas (bodies of knowledge) in the two contexts at encoding. The interaction we observed between "usefulness" and context-congruency is consistent with this idea: In our study, the items marked useful at encoding (i.e., items that could be meaningfully integrated into the active schema) were the ones that suffered when the schema active at encoding was not active at recall. Note that this interaction effect cannot be solely explained by deeper encoding of useful items (e.g., survival-related items, Soderstrom and McCabe, 2011) – depth of encoding can explain the main effect of usefulness but not the benefit of reinstating context. Having said this, our study did not directly manipulate schema-integration, so we need to temper our conclusions about the role of schema-integration in driving these effects.

Relating this to the literature more broadly, a potential reason why the seminal Godden and Baddeley (1975) study found such robust effects could be that their subjects were well-familiarized with the sequence of actions of underwater diving as well as those of being on land; consequently, they had "underwater" and "out of the water" schemas that they could use to scaffold knowledge at encoding. Our findings also resonate with prior work showing the importance of integrating items with context (Eich, 1985; Murnane et al., 1999). Eich (1985) showed a larger reinstatement effect in free recall when subjects integrated items into physical contexts (i.e., distinct rooms) by imagining them in the study environment. Similarly, Murnane et al. (1999) found larger context-change effects on recognition sensitivity when items could be integrated into a familiar situation (e.g., when words were shown on a picture of a blackboard in a classroom) vs. when they could not (e.g., when the context was defined by combinations of word color, background color, and location). Our usefulness-judgment results extend this idea by showing that simply having a meaningful context (i.e., one that activates an existing schema) is not sufficient to yield context-dependent memory; rather, subjects have to actually succeed in integrating the item into the encoding context (i.e., the item has to be judged to be "useful" in that context) to get an effect of context congruency for that item at recall.

Our study used VR in order to provide both perceptually and semantically rich experiences. The immersive nature of VR makes it a potentially useful tool for studying context-dependent memory (Dunsmoor et al., 2014; Reggente et al., 2018). However, the use of VR does not guarantee a context-dependent memory effect. For instance, a recent study that also used VR did not find the context-reinstatement effect (Wälti et al., 2019). In their study, Wälti and colleagues asked subjects to remember a list of words presented on a background image, following the study by Isarida and Isarida (2007) where contexts were manipulated using the background color for studied words. On each trial, a word was presented on a context image (i.e., a background color, a landscape picture, a virtual background, or background flickering) for 3 seconds, and the context was pseudorandomly selected on each trial such that the same context did not appear for more than three words in a row.

While we also used VR, our study significantly differs from the Wälti et al. (2019) study. Wälti and colleagues used VR to present backgrounds and did not allow direct interaction with these backgrounds; by contrast, we used highly interactive virtual environments to encourage activation of distinct schemas in the two environments. Moreover, our study used a surprise memory test to prevent participants from using other strategies that could potentially suppress processing of context information. Lastly, the frequency of context switches was higher in the Wälti study – there were at least seven context shifts during encoding. This context manipulation may have been ineffective because the rate of context switch was too rapid to match the human prior for a context change. For instance, it is unlikely that the room where we read emails changes as quickly as the switch of tabs on a browser or apps on a phone. In other words, it may be that contexts shift at hierarchically different timescales (Zacks et al., 2001b; Kurby and Zacks, 2008; Collins and Frank, 2013). If the two alternating contexts form a higher-level context in which there is an alternation between the two background images with certain transition probabilities, then what is designated as the “change” or “different context” condition may not actually involve a context change (since subjects continue with same higher-level mental context active in mind). Another implication of the idea that changes in the internal context representation drive memory effects (more so than changes in sensory input) is that asking participants to mentally reinstate the encoding context should reduce the size of the context-dependent memory effect (as in Smith, 1979; this is a promising direction for future work.

Lastly, in addition to the factors outlined above, we manipulated study-test delay: There were two delay conditions, one in which subjects were tested immediately after encoding and another after a 1-day delay. Our data show that the benefit of context reinstatement for the 1-day delay was numerically (but not significantly) larger than that of immediate recall. Increasing the delay might boost context-dependent memory effects by boosting the extent to which participants rely on the hippocampus (Chen et al., 2016) instead of PFC-mediated active maintenance of items (for a review, see Richmond and Zacks, 2017). Hippocampal codes are known to be highly context-dependent (Eichenbaum, 2004), so it follows that anything that increases reliance on the hippocampus should increase context-dependence. Future studies can use neuroimaging to test the role of hippocampal engagement in driving the context-dependent recall effects observed here.

In summary, we showed a context reinstatement effect using environments that were designed to activate pre-existing schemas (i.e., underwater and outer-space planet environments), and schema-consistent items were most likely to show context-dependent memory effects. These results suggest that integration of items into active schemas plays a key role in driving context-dependent recall effects. However, identifying the exact features of our paradigm that were necessary and sufficient for obtaining these effects will require additional research.

Chapter 3

The organization of episodic memory and latent cause inference

The contents of this chapter were published in Shin, Y. S., & DuBrow, S. (2020). Structuring Memory Through Inference-Based Event Segmentation. Topics in Cognitive Science https://doi.org/10.1111/tops.12505.

3.1 Introduction

Humans have a natural tendency to segment the continuous stream of incoming information we experience into discrete events, broadly defined as units of activity with an identifiable beginning and end (Zacks et al., 2007). People can detect the transitions between events, called event boundaries, while reading (Speer and Zacks, 2005) and watching films (Newtson, 1973), as well as during more naturalistic first-person experiences (Magliano et al., 2014). Event segmentation may occur spontaneously, as evidenced by longer processing times at boundaries in the absence of a segmentation task (e.g., Speer and Zacks, 2005; Hard et al., 2011). In addition, neuroimaging data has shown that the brain responds to boundaries during passive viewing (Zacks et al., 2001a; Speer et al., 2007; Ben-Yakov and Henson, 2018; Baldassano et al., 2017), representing information within an event more similarly than across events (Chen et al., 2017). Segmentation behavior tends to be consistent across people both behaviorally (Newtson, 1973; Jeunehomme and D'Argembeau, 2018; Zacks et al., 2001b) and neurally (Speer et al., 2003; Ben-Yakov and Henson, 2018; Baldassano et al., 2017), suggesting that event structure is construed in a systematic way. Moreover, segmentation has been shown to have important behavioral consequences such as enhancing memory for items encountered at boundaries (Swallow et al., 2009; Heusser et al., 2018; Rouhani et al., 2020) and warping time perception such that intervals with boundaries are estimated as longer in memory (Lositsky et al., 2016; Ezzyat and Davachi, 2014). Here, we first review the behavioral effects of event segmentation, then we review potential mechanisms of event segmentation, and finally we explore the role of inference as a framework for event cognition.

3.2 Why do we segment events?

Segmenting experience into distinct events has been shown to have extensive psychological consequences (see Table 3.1). In this section, we review three domains affected by event segmentation: 1) in-the-moment processing of *ongoing* experiences, 2) memory organization of *past* experiences, and 3) making decisions that best serve the current situation. The effects of event segmentation observed in these domains demonstrate a benefit for adaptive behavior.

3.2.1 Facilitation within ongoing event processing

When encountering incoming information, event segmentation can facilitate processing by increasing access to the event that is currently being experienced. One measure that has been used to assess facilitated processing is how long it takes to read narrative passages with and without event boundaries. Reading time increases have been observed in studies that explicitly signal time shifts (e.g., "an hour/day later" versus "a moment later", Zwaan, 1996; Speer and Zacks, 2005; "The next morning", Pettijohn and Radvansky, 2016) as well as protagonist changes (Rinck and Weber, 2003). These results suggest that narrative comprehension is facilitated within events versus across event boundaries, perhaps because within-event content is predicted by learned event schemas that involve sequences of states (Radvansky and Zacks, 2011; Franklin et al., 2019) or situation models (Zwaan and Radvansky, 1998). Zacks, Speer and Reynolds (2009) explicitly probed event segmentation and predictability and found that narrative passages rated low in predictability showed longer reading times and more identified boundaries. Moreover, consistent with a role for predictability in mediating event segmentation, Pettijohn and Radvansky (2016) showed that reading times do not slow down at shifts when they are foreshadowed (i.e., predicted). These data suggest that predictable content that belongs to the same event facilitates rapid comprehension.

Similar to the reading time data, when presented with still frames of action sequences, people tend to dwell longer on transition frames, suggesting increased processing demands at boundaries (Hard et al., 2011). Interestingly, however, predictability generally enhances this dwell time effect. That is, the dwell time difference between actions at boundaries and within-event actions increases as a function of experience (Hard et al., 2019; Kosie and Baldwin, 2019). This exaggeration may be driven both by reduced processing time within events as the predictability of actions increases, as well as by longer dwell times at boundaries as the anticipation of change triggers viewers to gather more information .

When processing ongoing streams of input, it may be helpful to have selective access to currently relevant information. Event segmentation can help prioritize information relevant to the currently active event, while making the information from previous events less accessible. This has been demonstrated in paradigms that interleave narrative reading or movie watching with recognition memory tests. When there is an event boundary between the encoding of the probe and the recognition test, people recognize the probe more slowly (Zwaan, 1996; Speer and Zacks, 2005) and less accurately (Speer and Zacks, 2005; Swallow et al., 2011). In a similar set of studies in which people experienced event boundaries by walking through doorways, items learned in a previous room became less accessible (Radvansky and Copeland, 2006; Radvansky et al., 2011). This suggests that when an event ends, the information that was learned within that event drops out of active working memory because it is no longer relevant. Updating an active event in working memory (i.e., event model) in this way is one of the principles of the Event Horizon framework (Radvansky and Zacks, 2017; Radvansky, 2012) and has beneficial effects for online processing of incoming information within the same event (e.g., reduced reading time and increased working memory access). Dropping previous information out of working memory may have the added benefit of preventing that irrelevant information from intruding on the current processing. This may also protect information in the previous event from retroactive interference, increasing the accuracy of later reconstruction (Gershman et al., 2014).

3.2.2 Memory organization

Segmenting events can also support the encoding and retrieval of episodic memories. One line of evidence that event segmentation helps episodic memory is that items that belong to the same event show mutually facilitated recognition memory. For instance, when people read multiple sentences, some of which share a location, they are faster to recognize the sentences that share a common location compared to sentences that do not (Radvansky and Zacks, 1991; Radvansky et al., 1998). This benefit may be driven by recent activation of a sentence that could prime retrieval of other episodes in the same event. One study that supports incidental retrieval of within-event items used a subset of a previously studied group of words as a target for an unrelated memory task, and showed that a lure from the same group was more likely to be falsely recognized than a lure from a different group (Hoskin et al., 2019). While this demonstrates that it can come at a cost at times, having segmented structure in memory can keep related information together, facilitating later retrieval.

Memory studies have also directly probed whether people are more likely to retrieve items that belong to the same event together. Analogous to the suppression of previous event information observed during ongoing processing, retrieving events from episodic memory that have a boundary between them can be more challenging than retrieving information from the same event. Zwaan (1996) used a cued recognition paradigm where a sentence that was followed by a time shift signal phrase served as a cue to facilitate recognition speed of the next target sentence. Event segmentation was manipulated by shift magnitudes (“a day/hour/moment later”). Consistent with the prediction that event segmentation enhances retrieval of items in the same event and diminishes retrieval of items in a different event, people were slower to recognize the target sentence after a larger shift. Using a similar time shift signal ("an hour later" vs "a moment later"), Ezzyat and Davachi (2011) asked people to recall what came after a cued sentence. In the large shift condition, recall performance was lower when a pre-shift sentence was used as a cue as compared with when a post-shift sentence served as a cue. However, a difference was not observed when the time shift was small. Complimenting the online predictability effects discussed in the previous section, these results suggest a mechanism by which retrieving an episode from long-term memory can cue the next episodes that occurred within the same event, guiding predictions of what will happen next.

Relatedly, studies that probe the temporal order of items that either belong to the same event or different events provide additional evidence for facilitated within-event retrieval. In DuBrow and Davachi (2013, 2014), subjects judged the relative recency of two items within or across boundaries that were created by switching stimulus categories and their associated tasks (e.g., male/female judgment for the face category; bigger/smaller than shoebox judgment for the object category). Recency judgments between two studied items were less accurate when the intervening sequence contained boundaries. To examine whether this performance drop was due to retrieval failure for the intervening items, recognition memory for those intervening items was tested immediately after (i.e., primed by) the order judgments. Consistent with the aforementioned studies that tested cued recall and cued recognition, when people made a correct order judgment, the speed at which they recognized the intervening items was faster when there was no event boundary, suggesting greater within-event access to item sequences. The within-event versus across-events difference in temporal order memory has also been shown in more recent studies that used perceptual boundaries (background color changes, Heusser et al., 2018) and spatial boundaries (room changes, Horner et al., 2016; turns in navigation, Brunec et al., 2018). These experiments show that temporal information is better preserved within an event than across events, potentially via better reconstruction of a study sequence.

Another way to test how individuals reconstruct sequences from memory is to examine transition probabilities in verbal recall. In one study, people were asked to recall items in the same order that they were studied. Accurate serial transitions between recalled items were found to be more common within than across category boundaries, providing additional support for the better reconstruction of a study sequence within an event (DuBrow and Davachi, 2016). Similarly, in unconstrained free recall studies, event-level clustering (Polyn et al., 2009) and a tendency towards more forward serial transitions within events compared to across boundaries (Heusser et al., 2018) have been observed. Since recall order was unconstrained, these results suggest that event-level organization may be a fundamental property of recall. That is, event structure may provide a scaffold for spontaneously recalling past experiences in their sequential order.

3.2.3 Adaptive decision making

Interestingly, this sequential recall closely resembles sequential reactivation in the hippocampus (Foster and Wilson, 2007), in which event structure has been observed while an experience unfolds. In particular, Gupta and colleagues (2012) showed that while rats navigate a maze, hippocampal activation reflects the segment of the environment that is currently being navigated (i.e., the event model), disproportionately representing paths ahead within the segment in the beginning and paths behind within the segment as they approach its end. This activation of currently relevant information has implications for decision making, where generalizing relevant past experiences can guide decisions for unknown possible futures. Indeed, an extensive literature on rodent navigation and decision making has shown that hippocampal activation of forward trajectories occurs *preceding* decisions about where to go next (Pfeiffer and Foster, 2013; Johnson and Redish, 2007; for reviews, see Ólafsdóttir et al., 2018 and Pezzulo et al., 2014). Similar neural reinstatement effects, both during rest (Schuck and Niv, 2019; Momennejad et al., 2018) and prospectively at decision times (Doll et al., 2015), have also been shown to influence subsequent decision-making in humans.

Although traditionally decision making has been viewed as relying on a representation of incrementally-learned value that is independent of episodic memory (Knowlton et al. 1996; cf. Poldrack et al. 2001), replay of past experiences at decision points suggests the potential contribution of episodic memories. Indeed, a growing literature has shown that adaptive decisions are influenced by episodic memory retrieval (Shohamy and Daw, 2015; Shadlen and Shohamy, 2016). For example, people are more likely to choose previously encountered items that were associated with high values than low values only when they remember such associations (Murty et al., 2016), and they may rely more on individual episodes over summary values for decision making following memory retrieval (Duncan et al., 2019). Linking the role of episodic memory and event segmentation, Bornstein and Norman (2017) showed that, when reminded of an image experienced in a previous event, people's decisions are biased by the summary of their experiences in that event, not just the specific experience associated with the reminded image. This result suggests that event segmentation can support the interaction between episodic memory and decision making, by guiding retrieval of past decision outcomes from previous events that most closely match the current situation.

Together, these studies suggest that organizing information into event structures can have important benefits for online processing and retrieval of relevant information that, in turn, can help guide adaptive decision making. Given the widespread effects of segmentation, it is crucial to appropriately segment our everyday experiences into events in a way that promotes later utilization. One of the major challenges in doing so is the inherent ambiguity of when an event starts and ends. Many of the studies reviewed showing the benefits for memory and decision making of event segmentation cannot address this challenge, as they imposed stark changes in spatial contexts, perceptual features, task sets, and so forth to manipulate segmentation (see Table 3.1). Experiments that use narratives to induce event boundaries are more ambiguous because there can be multiple changes along different dimensions (cf. event indexing theory; Zwaan et al., 1995a), but they still often contain signals for boundaries such as time shifts or scene changes. When event boundaries are not overtly signaled, how do we segment events? Below we review theories of event segmentation under ambiguity. We focus on the traditional prediction error account and our proposed framework that identifies boundaries based on changes in inference rather than extrinsic change.

3.3 Potential Mechanisms of event segmentation

3.3.1 Prediction error

The dominant account of event segmentation is that “prediction error”, the difference between one’s experience-based expectation and the currently observed outcome, signals the end of events and induces event boundaries (Zacks et al., 2007). To test how explicit predictions are related to event segmentation, Zacks and colleagues (2011) showed people movie clips of everyday activities, and occasionally paused the clip to ask them to predict what would happen in 5 seconds. When there was an intervening event boundary, as identified by independent observers, the prediction accuracy dropped, and there was greater activation in the substantia nigra, a region traditionally associated with dopaminergic responses to reward prediction errors. In line with this idea, a neural network model that implements perceptual prediction error as a gating signal to update event representations can identify simulated event boundaries

(Reynolds et al., 2007). However, there are major challenges regarding the precise relationship between prediction errors and event segmentation.

First, prediction errors may not always signal meaningful changes in event perception, particularly when the environment is uncertain (O'Reilly, 2013). That is, while it can be ideal to draw boundaries in order to discount the past when the underlying statistics of the world change abruptly, disregarding the broader environmental context would be suboptimal when the environment is noisy, as frequent high prediction errors would lead to over-segmentation. Instead, when prediction errors are frequent, boundaries should only be drawn following *unexpected* deviations (Dayan and Yu, 2006), as expected deviations are not reflective of meaningful changes in the underlying event structure. This can be accomplished by scaling down the degree to which prediction error updates expectations (i.e., the learning rate) for these expected deviations, such that the overall update in value would be small for expected deviations. Empirically, boundary-related memory effects that occur when changes are infrequent are absent when changes occur frequently (Siefke et al., 2019). This idea is further reflected in studies where people's learning rates dynamically adjust to the uncertainty of the environment, reducing the effects of prediction errors in noisy environments while enhancing them in stable environments, in which a relatively high prediction error indicates meaningful change (Pearce and Hall, 1980; Behrens et al., 2007; Nassar et al., 2010, 2012). This suggests that the magnitude of the update (i.e., prediction error × learning rate) may be a better indicator of event boundaries than prediction error alone, given the non-stationary nature of everyday experiences.

Second, prediction errors may not be necessary to create an event boundary. For example, in narrative reading, when an event shift is foreshadowed, readers still respond more slowly to the pre-shift memory probes even though they no longer experience surprise nor slow their reading (Pettijohn and Radvansky, 2016). This suggests that even expected change, when sufficiently meaningful, can drive event segmentation in memory. One demonstration of this is statistical learning (Saffran et al., 1996). In one study, Schapiro and colleagues (2013) found that people could identify boundaries in a series of stimuli based on their learned temporal transition statistics without ever experiencing prediction errors. In their experiments, most stimuli were exclusively followed by other stimuli within the same group, while some stimuli served as entry/exit points where a transition across groups was available. Notably, the individual transition probability from a group's exit point to a neighboring group's entry point was equal to the probability of each within-group transition, and thereby did not induce prediction errors (see also Richmond and Zacks, 2017, for discussion). After learning, participants successfully indicated the transitions between groups (i.e., the entry points) in the sequence of stimuli. This highlights the importance of segmenting experience even at predictable transitions between events. Indeed, predictable transitions may enhance segmentation effects by shifting attentional resources to boundaries .

As reviewed here, prediction error defined as the difference between one's expectation and the observed outcome may not be sufficient nor necessary for segmenting events. This calls for an alternative account of the mechanisms underlying event segmentation.

3.3.2 Inference of event types

No two events can be exactly the same even when they are from the same category, as at least one dimension, time, is always unique. For example, when items occur in two different instances of the same category (i.e., ABA structure) there is an intervening change in event type, and the two visits to A are idiosyncratic in terms of time. When comparing this scenario to a single continuous instance of an event (AAA structure), segmentation effects are observed despite comparing two items of the same type (DuBrow and Davachi, 2013, 2014). Radvansky et al. (2011) used a more naturalistic boundary manipulation with the same structure by having people walk around a series of rooms and then testing their memory. When returning to the same room after having been to a different room (ABA), their memory was worse than when they never left the room at all (AAA). Thus, staying in the same event instance (no event boundaries) has a memory advantage over merely having shared context (room A). This suggests that boundaries affect memory above-and-beyond the effects of changing context.

Time-sensitive and idiosyncratic *event instances* do interact with the structure of knowledge formed by multiple encounters with similar situations. In Radvansky et al. (2011), people also performed better when they returned to the same room (ABA) compared to when they went to another room (ABC). Thus, the *event type* that represents a class of experiences (called event schemata in Event Segmentation Theory, Zacks et al., 2007) is a key factor in understanding segmentation (for reviews, see Zacks et al., 2007 and Radvansky and Zacks, 2011). For example, in the statistical learning study described in the previous section, stimuli that tended to be experienced in close temporal proximity became discrete clusters (event types) in memory (Schapiro et al., 2013). When revisited, these clusters could be recognized even though different paths were experienced across learning instances due to the probabilistic structure. These studies raise the possibility that event types can be a useful basis for promoting generalization of learning within type, allowing us to extrapolate our previous experiences in an event to a new instance.

Event types can also stabilize event segmentation hierarchically such that, in parsing events, low-level perceptual changes that are not relevant to the event type carry less weight than changes along dimensions that are pertinent to the event type. In a neural network application of this idea, the REPRISE model stabilizes low-level perceptual and motor information to be consistent with the current event type (called a “context vector”) when executing event-level control (Butz et al., 2019).

Empirically, hierarchical structures in event segmentation are consistently observed in behavioral (Zacks et al., 2001b) and brain imaging studies (Hasson et al., 2008; Lerner et al., 2011; Baldassano et al., 2017), supporting the idea that event types are utilized in segmenting experience.

It is important to note that these event types do not exist in isolation and are continuously created and updated based on event instances. How do we dynamically update and create new event types? To inform this, we turn to category learning, where a similar set of challenges exists. Categories, like event types, help facilitate understanding of their individual members through generalization of category properties. However, as in event perception, we do not know how many categories exist in the world, and we need to update the existing categories' summary statistics based on their members, while being open to creating a completely new category. The rational model of categorization addresses these challenges by proposing that a stimulus is more likely to be a member of a prolific category with many members, and yet a new category can be inferred at any point when the existing categories are not a good match (Anderson, 1991; also called the Chinese Restaurant Process (CRP), Aldous, 1985). In discovering clusters such as categories from observations, this rich-gets-richer process reins in the tendency to create too many clusters and thus keeps the model simple while still allowing for new cluster formation. The adaptable nature of the model aids predictions within categories by extrapolating features and functions within the categories, whether or not an explicit label was given to a new category (Anderson, 1991). Thanks to its flexibility, this model can be generalized to two popular models of categorization where instances of a category can either 1) be lumped together such that one summary value can account for the category (prototype model; Reed, 1972) or 2) be perfectly preserved to be later compared with other instances (exemplar model; Medin and Schaffer, 1978; Nosofsky, 1986), by varying a single parameter that governs creation of a new category (Sanborn et al., 2010). The rational model provides a unified framework that successfully predicts categorization behavior in humans across domains (Anderson, 1991; Sanborn et al., 2010).

The model is particularly useful in situations where there is a latent variable (i.e., a cluster of experiences such as a category or event type) that needs to be discovered to properly generalize across experiences. For example, the model can explain how people effectively generalize between tasks by forming clusters of task sets (Collins and Frank, 2013). In reinforcement learning, the model can explain how the latent variable can guide reward predictions, and how a new latent variable can be inferred when existing ones do not predict the outcome (Gershman et al., 2010). This model performs better than classic reinforcement learning models in explaining how compound reward cues are flexibly represented (Courville et al., 2005; Soto et al., 2014) and why conditioned responses come back after extinction (Gershman et al., 2010). In event perception, the latent variable would correspond to event types. That is, we can identify an event type from a time-specific event instance, updating the type based on the instance, or create a new event type when an instance does not fit with any of the existing event types' properties. The latent variable model would also predict that, due to its rich-gets-richer property, we are more likely to infer an event type that we encounter often (i.e., high prior probability) when it sufficiently explains the current episode (i.e., high likelihood), rather than inferring a new event type *de novo*. In this framework, event segmentation is likely to occur when the distribution over inferred event types diverges from the distribution at the previous time point. Event segmentation based on distributional changes can explain how experiences are clustered into units even when there is no prediction error between experiences, as in statistical learning-based clustering described above.

A demonstration of this inference process is depicted in Figure 3.1. As we watch a movie, we learn that the character Julie recently lost her husband Patrice, a famous composer, in an accident. Later, Julie discovers that Patrice's assistant Olivier is

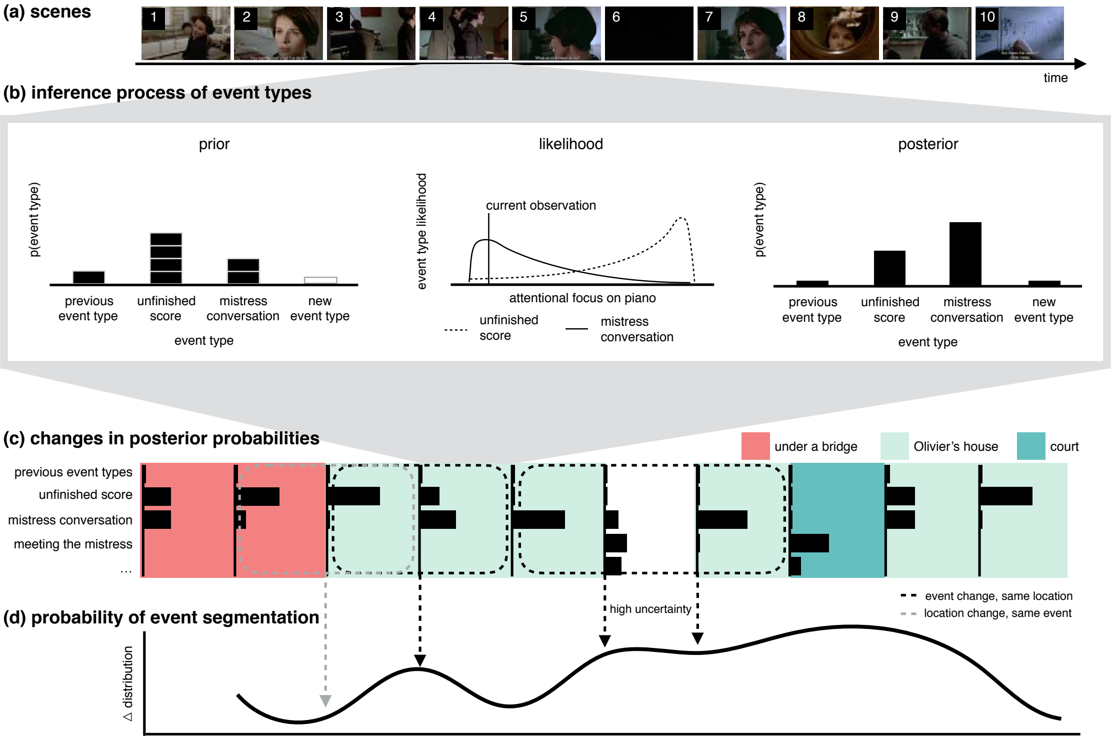

Figure 3.1: An example of inference-based segmentation
(a) scenes from Krzysztof Kieslowski's movie *Three colors: Blue.* In this movie, Julie, a recent widow of a famous composer, tries to finish her husband's score and tracks down his mistress. **(b) inference process of event types** *left*: Prior probability of each event type reflects the popularity of each event type (indicated by number of boxes constituting the bars). Note that there is a small chance of creating a new event type that was never observed yet. *Center*: The likelihood of event type "mistress conversation" is higher than the likelihood of "unfinished score" given the current observation (low attention on the piano). *Right*: Combining the prior distribution and likelihoods gives the current posterior probability distribution over event types. Here, posterior probability is higher for the "mistress conversation" than the "unfinished score" event type, although the prior was higher for "unfinished score". **(c) changes in posterior probabilities** The posterior distribution over event types changes as observations change. Note that changes in the location (denoted by the background color) are not always predictive of posterior distribution changes. **(d) probability of event segmentation** Events are more likely to be segmented when changes in the probability distribution from the previous time point are large. Again, event segmentation (black dotted lines) and location change (grey dotted lines) do not always correspond.

trying to complete one of Patrice's unfinished scores *and* that Patrice had a mistress. When we see Julie approaching Olivier (Figure 3.1.a-1), we do not know whether she will confront him about the score or the mistress, and the probabilities of the two event types are uniformly distributed (Figure 3.1.c leftmost panel). When Julie and Olivier have a conversation about the unfinished score (Figure 3.1.a-2) and then the movie jumps to Olivier playing the song (Figure 3.1.a-3), the same "unfinished score" event type is active despite the perceptual input and spatial context having changed (denoted by the change in background color in Figure 3.1.c). Note that the probability of event segmentation is low despite the location change (denoted by a grey dotted line in Figure 3.1.c and d). Conversely, event segmentation can happen without large perceptual changes (denoted by black dotted lines in Figure 3.1.c and d). In *scene 4*, the prior for "unfinished score" is higher than "mistress conversation" because of the greater number of previous instances of the "unfinished score" event type (denoted by the number of boxes in (Figure 3.1.b left) and the model's property that favors prolific event types. However, despite occurring at the same location, as Julie diverts her attention away from the piano to ask Olivier whether he knew about the mistress, the likelihood of the "unfinished score" event decreases while the likelihood of the "mistress conversation" increases (Figure 3.1.b center). Thus, the updated posterior probability shifts towards "mistress conversation" (Figure 3.1.b right). In the proposed framework, the divergence between the updated probability distribution and that of the previous time point increases the probability of event segmentation (Figure 3.1.d).

Inference-based segmentation can be useful when there is high uncertainty about the upcoming event type, as it allows an event to end without introducing a new event, thereby protecting the event that just ended from further interference. In the example, when Olivier asks what Julie wants to do about the mistress (Figure 3.1.a-5), the screen fades out (Figure 3.1.a-6). With no visual input, we can still make predictions about what will happen in the next scene. It turns out that the next scene is the continuation of *scene 5* where Julie answers Olivier that she wants to meet the mistress (Figure 3.1.a-7). Note, the probability distribution at *scene 7* is similar to the one in *scene 5*, but event segmentation is still likely to occur between *scenes 5* and *7* because the prediction at *scene 6* became highly uncertain and thus the probability distributions diverged.

One key aspect of the framework is that the inference process makes use of previously created event types. For example, after meeting the mistress, we again see Julie visiting Olivier (Figure 3.1.a-9), and are again unsure about what the topic of their conversation will be. Thus, the probabilities of the two event types are uniformly distributed. When we see that they start working on Patrice's score together (Figure 3.1.a-10), instead of creating a whole new event type, we can reactivate the "unfinished score" event type and update the event with new information. Again, notice that the probability distributions for *scenes 3* and *10* are similar, but segmentation has occurred between those two event instances. Overall, this example illustrates the inference process and three key features of the model: 1) event segmentation occurs when the inferred event rather than observed features change, 2) common event types are more likely to be inferred, and 3) previous event types can be revisited and updated, enabling generalization.

3.3.3 Support for the latent variable account

Beyond online inference during event perception, the latent variable account also makes specific predictions about memory retrieval related to its cluster organization of individual episodes. As an analogy to memory retrieval, imagine looking back at some photo albums of a recent trip to Hawaii during which you went surfing many times and hiking just once (Fig. 2). You may look at the two photo albums "Surfing in Hawaii" and "Hiking in Hawaii" and summarize your Hawaii trip as "surfing and hiking." That is, summarizing experiences by sampling at the cluster (photo album) level leads to an over-representation of rare episodes relative to how often they were actually experienced. In this example, cluster-level sampling will bias the memory towards equal weighting of surfing and hiking. By contrast, episodic sampling of the individual images will accurately represent their relative frequency and reflect that surfing was the main activity.

Relying on latent structure for summarizing experiences can also explain systematic memory distortions towards gist–the summary statistics of the latent variable. For instance, in a study where the organizing structure was imposed by item categories (e.g., lamp), color memory for individual members of a category was distorted toward the center of the color distribution of its category (Brady et al., 2018). At its extreme, the latent variable can even create false memories such that words that are not a part of the studied list yet exist at the conceptual/semantic center of the list are falsely recalled as one of the studied items (Deese, 1959; Roediger and McDermott, 1995). These types of gist biases can be useful in summarizing experiences, albeit at the cost of accuracy for each individual episode.

While the latent variable model can summarize experiences at the cluster level, this does not preclude sampling individual episodes during retrieval. Indeed, episodic sampling from clusters makes additional predictions for memory biases. Returning to the photo album analogy, imagine choosing a few photos from each photo album. Any given photo from an album with fewer photos would be more likely to be chosen. In a similar sense, when searching memory for past experiences by going through latent variables (events) and sampling from each of them (cluster-level sampling; Figure 3.2 top), the episodes that have fewer companions in a cluster are more likely to be sampled, and thus are more likely to be retrieved. This idea is consistent with work by Alves and colleagues (2015) where people were found to be better at recognizing words that have fewer close conceptual neighbors in the same studied list. Sampling

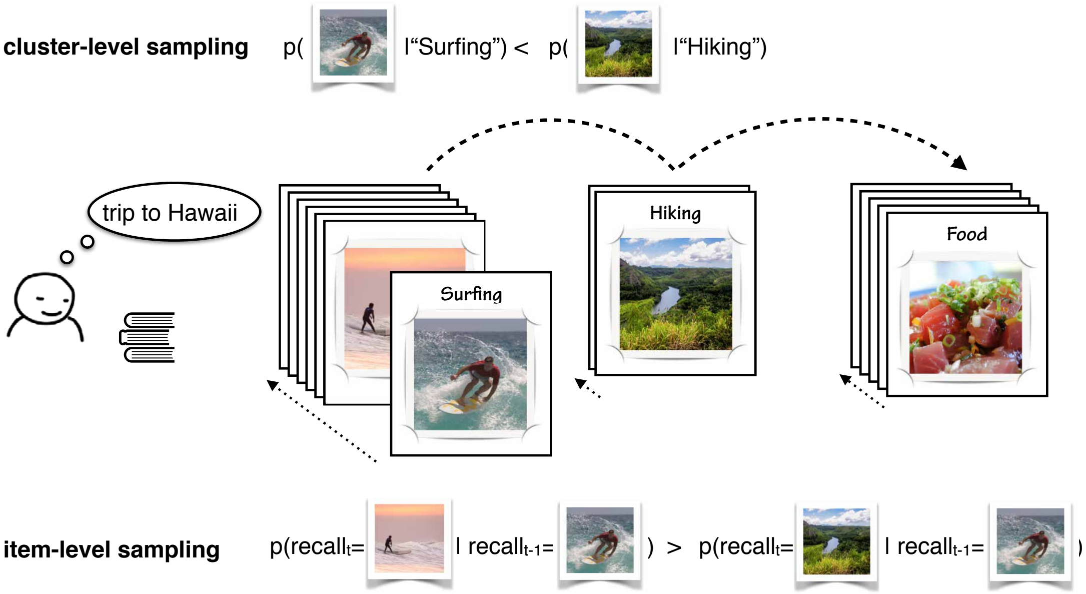
cluster-level sampling
p(|"Surfing") < p(|"Hiking")
trip to Hawaii
Surfing
Hiking
Food
item-level sampling
p(recall_t= | recall_t-1=) > p(recall_t= | recall_t-1=)

Figure 3.2: Schematic for cluster-level and item-level sampling
As an analogy to memory retrieval, imagine looking back at some photo albums of a recent trip to Hawaii. If you summarize the trip, you may look at a few photos from each album (cluster-level sampling; top dashed arrow). In this case, photos from a thinner album (e.g., "Hiking") would be more likely to be picked than ones from a thicker album (e.g., "Surfing"), gaining prominence in the overall summary. On the other hand, you may want to retrieve detailed experiences, looking at each photo (item-level sampling; bottom dotted arrows). In this case, a photo will be more likely to be followed by another photo from the same album (e.g., "Surfing") than a different album (e.g., "Hiking").

in this way can also account for memory biases like the von Restorff effect in which distinctive items are better remembered (von Restorff, 1933; Hunt, 1995) and the cue overload effect in which memory is better for items whose retrieval cue has fewer associated items (Watkins and Watkins, 1976).

What determines how many episodes each cluster contains when the clusters need to be dynamically inferred? One answer is extreme events that deviate from the average of previously existing clusters in terms of the relevant feature dimensions (for example, a very loud noise that is not commonly experienced). During inference, the latent variable model is more likely to create an entirely new cluster for an extreme event. Since any given episode in a small cluster is more likely to be sampled, cluster-level sampling can account for memory and decision biases towards extreme episodes. For example, when people directly experience risky outcomes, they become risk seeking for gains and risk averse for losses (Hertwig et al., 2004; Ludvig and Spetch, 2011). These decision biases are mediated by a memory bias in which extreme outcomes are remembered better and judged to have occurred more frequently (Madan et al., 2014, 2017). Similarly, episodes are better remembered when they elicit high reward prediction errors, regardless of their valence (Rouhani et al., 2018). Recent work by Lieder and colleagues (2014; 2018) also demonstrates that extreme events are more likely to be sampled from memory, although they propose a different sampling mechanism.

When trying to search for specific episodes, it would be useful to skim over a range of events (e.g., skimming the cover of photo albums) and go deeper only once we find the relevant event (e.g., open a photo album and look through the album). Supporting this idea, neural replay patterns measured with magnetoencephalography (MEG) have been shown to skip between events, reinstating only the entry points of events in sequence (Michelmann et al., 2019). When retrieval continues (e.g., choosing photos one after another), however, episodic samples would be more likely to transition within events than across events (i.e., item-level sampling; Figure 3.2 bottom). That is, when we are motivated to retrieve details of episodic memories, rather than providing a quick summary of the entire experiences, it is easier to hold onto one album and to go through individual photos within an album than going back and forth between albums. This idea is supported by empirical data that show transitions within events are more likely than across events in free recall (Heusser et al., 2018) and serial recall (DuBrow and Davachi, 2016). The latent variable model would explain such recall behavior by modulating recall probability according to the similarity between posterior distributions over latent variables. Linking the latent variable model to episodic memory, Socher and colleagues (2009) showed that a variant of the model can predict within-event transitions in human recall behavior. In their variant, the probability of recalling a specific word was determined based on the mixture of the latent topic structure (semantic context) and the temporal adjacency (episodic context) active at a given time. This model could predict recall transitions better than models with purely semantic or episodic context, suggesting that both conceptual and temporal structures are critical features of clusters in memory.

3.4 Future directions

Theories of episodic memory organization provide additional insight into how temporal information may play a role in structuring events. For instance, the Temporal Context Model (Howard and Kahana, 2002) and the Context Maintenance and Retrieval model (Polyn et al., 2009) have emphasized how storing a separate temporal representation may provide a scaffold for organizing memories. While these models have been highly successful in predicting memory recall, the way in which time interacts with ongoing event encoding and segmentation needs further investigation. Rather than having to store an independent representation of temporal information

(Socher et al., 2009), a nonparametric Bayesian model in which time is incorporated into the process of inferring latent variables could be more parsimonious. That is, the model could be sensitive to the recency of events without having to separately track time, as the probability of a previous event would decay over time since it was last active. In addition, by assuming that recently encountered event types have a higher chance of producing the current observation, the model can provide temporal stability in the inference process. One candidate model is a latent variable model that utilizes distance in the prior probability of events, called the distance-dependent Chinese Restaurant Process (ddCRP; Blei and Frazier, 2011). Instead of relying on cluster popularities as the standard rational model of categorization and Chinese Restaurant Process do, the ddCRP prior can assign probabilities according to the temporal distance between previous and current observations. Similarly, a simpler variant of this model in which a currently active cluster gets an extra boost, called a sticky Chinese Restaurant Process (Fox et al., 2011), has been used to account for stable event perception (Gershman et al., 2014; Franklin et al., 2019). Another intriguing direction for event segmentation research is incorporating time into the ddCRP hierarchically (cf. Ghosh et al., 2011) for event types to perform a stabilizing function that can improve generalization across instances.

The exact mechanisms by which event types are inferred and how the inferred event type interacts with event segmentation need further exploration. One possibility is that events are segmented when there is a change in the distribution over inferred event types (Figure 3.1). This would explain the event segmentation effects observed when boundaries are imposed by time shift signals (e.g., “a while later”) before the next event begins. That is, such phrases or other signals that the previous event has ended and is no longer relevant (e.g., the fade-to-black sixth scene in Figure 3.1.a) would increase uncertainty over event types, thereby changing the probability distribution and inducing an event boundary. A conceptual parallel can be found in a neural network model called the Connectionist Temporal Classification model (Graves et al., 2006). In this model, boundaries can be created at moments of high uncertainty (e.g., silence) where the likelihoods of any existing labels (event types) are low. This approach of treating high uncertainty as a non-labeled state increases flexibility in terms of how long an event lasts by allowing an event to end before the next one begins. These models differ from Event Segmentation Theory, which assumes that a new event begins as soon as the previous event ends (Kurby and Zacks, 2008). Experiments that test specific hypotheses (for example, event boundaries will occur at the offset of an event instance as well as at the onset) based on the proposed framework would provide further insight as to how events are segmented from experiences, being guided by, and guiding, predictions.

There are remaining questions regarding how transitions between event types are learned and represented. In its current form, the latent variable framework does not directly address how transition probabilities would be incorporated in event segmentation. Neural network approaches have begun to investigate how transition models between low-level event types (behavioral primitives) can be learned such that the history of previous event transitions would inform the subsequent prediction at an event boundary (Gumbsch et al., 2019). Another important question pertains to the hierarchical structure of event types. That is, when one event type is repeatedly followed by another, would those two event types continue to be recognized as distinct or would they ultimately be merged into a single, more complex event type? In either case, the mechanisms by which transition probabilities (within and across events) are represented and utilized in event cognition will need further examination.

3.5 Conclusion

A large body of work now suggests that event segmentation is a fundamental process that emerges naturally and is remarkably consistent across individuals. Research on how segmentation influences memory demonstrates its adaptive utility in increasing access to relevant information and reducing interference. In order to better understand the cognitive operations that support event segmentation, we must examine patterns of behavior when the answer is not clear (i.e., during ambiguous transitions) and model the internal processes that infer change based on ambiguous input. In particular, we propose that latent variable inference provides a useful framework for characterizing how we identify when an event is no longer relevant and select among alternatives. This framework accounts for existing data on the consequences of event segmentation for online processing, memory, and decision making, and generates new predictions that can guide future research and model development.

In the next chapter, I test the prediction of cluster-level sampling discussed in this chapter, by modeling the inference of a latent variable based on a set of observations. I investigate how cluster-level sampling can be the basis of summarizing experiences and lead to biases in the summary.

Table 3.1: Behavioral effects of event segmentation by boundary manipulations

Boundary manipulation	**Study**	**Behavioral effect(s)**
		Reading time
Narrative changes	Zwaan et al. (1995b)	Temporal and causal change > no change
	Zwaan et al. (1998)	More situational change > less
	Rinck and Weber (2003)	Temporal and protagonist change > no change
	Zacks et al. (2009)	More situational change > less
	Radvansky and Copeland (2010)	Temporal change > no change
	McNerney et al. (2011)	Causal and character change > no change > spatial and temporal change
	Pettijohn and Radvansky (2016)	Unexpected change > expected or no change
Narrative time changes	Zwaan (1996)	Temporal change > no change
	Speer and Zacks (2005)	Temporal change > no change
		Memory access
Narrative time changes	Zwaan (1996)	Within > across recognition (online and delayed)
	Speer and Zacks (2005)	Within > across recognition (online)
	Ezzyat and Davachi (2011)	Within > across cued recall
Activity change (video)	Swallow et al. (2009)	Within > across nonboundary recognition (online); Across > within boundary recognition (online)
	Swallow et al. (2011)	Within > across nonboundary recognition (online); Across > within boundary recognition (online)
Task and category change	DuBrow and Davachi (2013)	Within > across order memory and serial recall
	DuBrow and Davachi (2014)	Within > across order memory
	DuBrow and Davachi (2016)	Within > across serial recall
Virtual room change	Horner et al. (2016)	Within > across sequence recognition
Background color change	Heusser et al. (2018)	Within > across order memory
Turns in virtual navigation	Brunec et al. (2018)	Within > across order memory; Across > within duration discrimination
		Spontaneous clustering
Narrative changes	Zwaan et al. (1995b)	Within > across verb clustering
Task change	Polyn et al. (2009)	Within > across recall transitions
Background color change	Heusser et al. (2018)	Within > across recall transitions
		Prediction
Narrative changes	Zacks et al. (2009)	Within > across perceived predictability
	Pettijohn and Radvansky (2016)	Within > across expectedness ratings
Activity change (video)	Zacks et al. (2011)	Within > across prediction accuracy

Chapter 4

Emergence of evaluation biases from latent cause inference

The contents of this chapter were published in Shin, Y., & Niv, Y. (2020, April 10). Biased evaluations emerge from inferring hidden causes.

4.1 Introduction

In impression formation, negative information tends to have a stronger impact than positive information. Thus, the overall impression formed after positive and negative information is more negative than the algebraic sum of valences of individual experiences (Rozin and Royzman, 2001). This phenomenon, called negativity bias, has been extensively demonstrated.

One potential source of the asymmetry in impression formation is the low frequency of negative events. As they are rare, negative events are more surprising and grab more attention. Supporting this idea, participants spend more time looking at negative descriptions of a target person than they do positive descriptions, and weighting valences of descriptions by looking time predicts the resulting evaluation bias (Fiske, 1980). Indeed, reinforcement-learning models suggest that when tracking a cue's value, surprising outcomes update the value more strongly, resulting in greater contributions of surprising—in this case, negative—events to the overall evaluation (Pearce and Hall, 1980).

Diagnosticity of information is another important aspect in biases. For instance, an intelligent person can occasionally behave unintelligently, but the opposite is less likely. As intelligent behaviors are more diagnostic of intelligence than unintelligent behaviors are, these diagnostic behaviors carry more weight and dominate evaluation even when participants observe evidence of unintelligence more frequently (Rozin and Royzman, 2001; Mende-Siedlecki et al., 2013). In social settings, in contrast, positive behaviors are often the default (Alves et al., 2017a), making negative events more diagnostic than positive ones. Therefore, negative information could weigh heavily in such situations. Work on conceptual similarities is also in line with this idea: pairwise similarity between negative words is judged to be lower than between positive words, meaning that the distribution of negative concepts is sparser than positive concepts' distribution (Unkelbach et al., 2008; Alves et al., 2015).

If post-hoc impression judgments could rely on perfect encoding and retrieval of memory, evaluations would accurately reflect experiences without distortions. However, we cannot remember every single encounter with every person we ever meet. Therefore, we might lump together different similar experiences, clustering them in memory for future use. The high variance in negative concepts could lead to less clustering of those concepts in memory, leaving each individual observation more diagnostic of the valence of its cluster. Positive concepts, on the other hand, may form a large cluster of positivity where each individual piece of information is less diagnostic of the overall "gist" of the cluster. Supporting the idea that positive-valence concepts are clustered together more than are negative-valence concepts, processing of a positive word is faster if it follows another positive word, whereas this is not the case for negative words (Unkelbach et al., 2008). This could explain how the sparsity of negative events, given by rarity and variance, seems to contribute to the negativity bias (Alves et al., 2018).

"Latent causes" – hidden causal structures that are assumed to generate a set of observable events, can be a meaningful basis for summarizing experiences (Courville et al., 2005; Gershman et al., 2010; Gershman and Niv, 2010). For instance, when we believe that a single underlying reason (e.g., a person wanting a future favor from us) is causing ten separate events that we observe (e.g., friendly encounters with the person), we can keep one summary of those ten events instead of trying to remember each one of them. However, if we believe that two separate underlying causes (e.g., wanting a future favor from us and being socially anxious) generated five events each (e.g., friendly encounters with the person and socially inadequate behaviors of the person), we might keep two summaries. We may also care about generalizing across people as a group, for example, for forming expectations about the norms of people in different countries we may travel to. Relying on latent causes is normative when generalizing past experiences to a new situation, as we can utilize what we learned from the past events that are caused by the latent cause we expect is active now, but not the ones caused by other causes.

This approach, although rational, can also be a source of biases because we do not know the true latent cause of each event. For example, we may infer distinctive latent causes from seemingly different events even if, in truth, the events share one cause. For instance, the sparsity of negative events may lead to inference of many distinctive latent causes, while positive events may be attributed to a small number of causes (Figure 4.1A). If we then make our overall impression at the level of these latent causes—for instance by averaging over the valence of all the causes associated with the person or group of people we are forming an impression about—the estimation will be biased toward the sparse area where there are a larger number of inferred latent causes. Here, we hypothesize that the negativity bias emerges from the combination of normative segmentation of information into causes based on similarity and incorrectly weighted averaging over latent causes.

To test this hypothesis, we manipulated the sparsity of event distributions such that either the below-average values or above-average values were sparser than the other (Figure 4.1B), and asked human participants to observe a sequence of events drawn from either type of the distributions, described as donations (Experiment 1A, Experiment 1B, Experiment 2A), sales (Experiment 2B), or rewards (Experiment 2C). We then asked them to estimate the average value of the observed events, and compared estimations between sparsity conditions, which all had the same true average. If biases arise from inferring underlying causes such as intentions behind behaviors, we would expect that the estimated average of a sequence consisting of sparse below-average events would be lower than the estimated average of a sparse above-average event sequence.

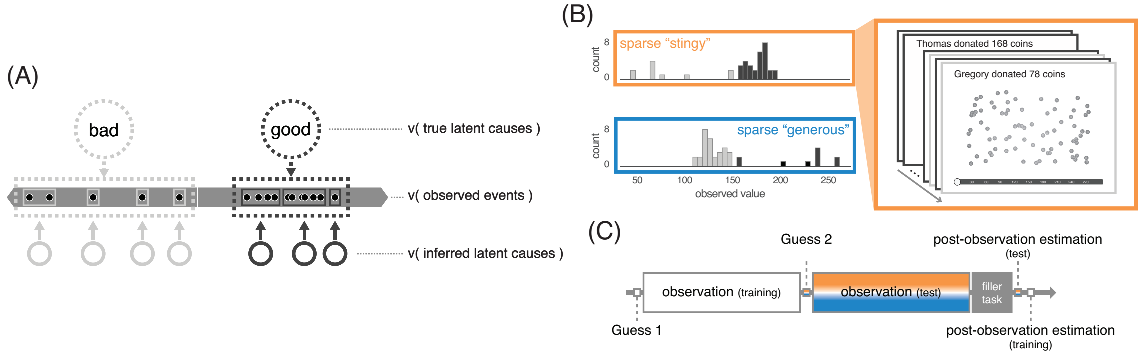

Figure 4.1: (A) Hypothetical latent structure of positive and negative events. Events (dots in middle) are generated from two true latent causes (dashed circles, top). The observer infers these latent causes (solid circles, bottom) from the observed events based on the similarity of events to each other. If events generated from the "bad" latent cause are sparse, the observer may infer many distinctive latent causes, each accounting only for a few observations. However, a small number of causes may be inferred to account for the many, similar, good events. (B) Experimental design for the sparse "stingy" donor and sparse "generous" donor conditions. The donation amount was drawn from "stingy" (below the mean; light gray) and "generous" (above the mean; dark gray) distributions. In the sparse "stingy" condition, most donations were drawn from a homogeneous "generous" distribution, while few, variable, donations were drawn from a "stingy" distribution. The distribution was flipped in the sparse "generous" condition. In both conditions, the true mean of donation amounts was 150. The subject observed the number of coins donated on each trial, marking the amount on a slider bar to register it. (C) Experimental Procedure. After subjects first guessed the general donation amount (Guess 1), we showed them a symmetric training sequence (with no sparsity) to adjust their expectations. They then guessed the general donation amount again (Guess 2) and observed a test sequence according to their experimental condition. After performing a filler task where they clicked a series of stimuli that appeared on random locations, subjects estimated the overall donation amount for the test and training sequences (post-observation estimations).

4.2 Experiment 1A

In Experiment 1A, we manipulated the sparsity of the distribution by variance and rarity in two sparsity conditions: the sparse ”stingy” and the sparse ”generous” conditions. We predicted that the estimation of average amount will be biased towards the sparse area, such that the estimates when below-average donations are sparse (the sparse ”stingy” condition) are below the estimates when above-average donations are sparse (the sparse ”generous” condition).

To test whether the biases can be predicted by an inference process, we compared human participants’ biases to simulated estimate biases from an approximate Bayesian inference model that inferred the latent structure using an infinite capacity prior called Chinese Restaurant Process (CRP). CRP assigned each donor to a latent cause, without predetermining the number of latent causes. That is, if a latent cause already generated many donors, the prior probability that this popular cause would generate the next donor is higher (“rich-gets-richer” property), and there is always a small chance that a completely new latent cause would produce the next donor, allowing flexibility in creating any number of causes. In combination with this prior probability, each latent cause generates donors with similar donation amounts.

We predicted that, as more distinctive latent causes will be inferred from sparse area, while a small number of causes may be inferred from dense area, events from the sparse area will have higher impacts than events from the dense area in the average estimate regardless of normalized valence (i.e., above- or below-average).

4.2.1 Methods

Participants

Eighty-three participants were recruited using Amazon Mechanical Turk (MTurk). The Princeton University Institutional Review Board approved the experiment, and we obtained informed consent online before participants began the task.

Participants were excluded when they did not pass the following criteria: (1) not completing the task to the end, (2) failing to answer correctly on attentional checks, (3) responding too slowly (> 60s) on any of the observation trials (4) not adjusting general expectation properly after the training sequence and making a guess with a value that they never observed during the training trials. This filtering was done to ensure that participants attended to every observation, as our prediction is based on the particular set of values of the events. Seventy-six subjects passed the criteria and were included in the following analyses (the sparse ”stingy” condition N = 34; the sparse ”generous” condition N = 42).

Task and Procedure

Participants were told that they were visiting different schools for fundraiser events and their job was to log the donation amounts. They observed two sequences of 40 donors making coin donations, where each donor made a single donation (Figure 4.1C). Each sequence was presented as a group of people who attended the same college. The first sequence served as a training sequence to adjust participants' prior beliefs about donations. After reporting their general expectation for donation (Guess 1) on a scale of 1 to 300 coins, participants observed a training sequence presented as donations from one college (“Brookview University”).

On each trial, coins were dropped on the screen with a prompt indicating the donor and the amount (e.g., ”Bradley donated 148 coins.”). Participants logged the amount either on a slider bar or in a text box next to the slider. The two response methods were yoked such that moving the slider would show the number in the text box and putting a number in the text box would move the slider to the number. The trial could proceed only when the response exactly matched the prompted amount.

For the training sequence, donation amounts were equally distributed above and below the mean donation (150 coins). To ensure that the training sequence indeed adjusted participants' prior beliefs close to the true mean, following this sequence we again asked participants to report their general expectation for donations (Guess 2).

We then showed a test sequence for which the true mean was the same as the mean of the training sequence, but the sparsity of below-mean and above-mean donors was manipulated between subjects. The test sequence was presented as donations from another college (“Cedar Springs University”). Participants in the sparse “stingy” condition ($N = 34$) observed a sequence where the below-mean ("stingy") donor distribution was sparser than the above-mean ("generous") donor distribution. In this condition, there were fewer stingy donors with higher variance in donation amount (10 "stingy" donors, $M = 79.7$, $SD = 35.86$) than generous donors (30 "generous" donors; $M = 173.73$; $SD = 10.28$). The amounts donated by the fewer "stingy" donors were more variable than the amounts donated by the many "generous" donors, to maintain the overall mean (see Figure 4.1B). Participants in the sparse "generous" condition ($N = 42$) observed a sequence whose donation values were flipped such that there are fewer and more variable "generous" donors (10 "generous" donors $M = 220.3$, $SD = 35.86$; 30 "stingy" donors, $M = 126.26$; $SD = 10.28$). After observing the test sequence, participants performed a filler task where they completed a questionnaire regarding their beliefs about human traits (Essentialism scale; Levy et al., 1998; Bastian and Haslam, 2006). Following the filler task, a surprise test asked them to estimate the average donation of the test sequence (post-observation estimate). This was followed by a test on the average donation of the training sequence.

Latent-cause Inference model

Each event sample was sequentially introduced to a Bayesian inference model with an infinite capacity Chinese Restaurant Process (CRP) prior (Aldous, 1985). In this model, before observing any behavior, an observer has prior beliefs about the target group's stable latent causes:

$$p(Z = k) = \begin{cases} \dfrac{n_k}{\sum\limits_{k=1}^{K} n_k + \alpha} & \text{for } k \leq K \\ \dfrac{\alpha}{\sum\limits_{k=1}^{K} n_k + \alpha} & \text{for } k > K \end{cases} \qquad \text{(Equation 4.1)}$$

where Z is a variable denoting the latent cause of the next observation, k indexes latent causes that ranges from 1 to K, n_k is the number of observations already assigned to latent cause k, and the concentration parameter α determines the prior tendency to assume new latent causes. This prior formalizes the idea that a prolific latent cause is more likely to generate future events (Equation 4.1 top case), and the total number of latent causes is unbounded and can grow with the number of observations (Equation 4.1 bottom case).

After observing an event, the likelihood that the current event x_t was generated from latent cause k is estimated by marginalizing over all ”consequential regions (Shepard, 1987)” h' that encompass the past events $\{x_i\}_k$ generated by cause k:

$$p(x_t \in k|\{x_i\}_k) = \sum_{h' \in H} p(x_t \in k|\{x_i\}_k, h')p(h'|\{x_i\}_k) \qquad \text{(Equation 4.2)}$$

The posterior probability $p(h'|\{x_i\}_k)$ in the righthand side of Equation 4.2 is calculated as:

$$p(h'|\{x_i\}_k) = \frac{p(\{x_i\}_k|h')p(h')}{\sum_{h \in H} p(\{x_i\}_k|h)p(h)} \qquad \text{(Equation 4.3)}$$

where the prior $p(h')$ follows an Erlang distribution (Shepard, 1987) with a size prior set to the range of training sequence events, and the likelihood $p(\{x_i\}_k|h')$ is the product of the likelihood of events that are sampled from consequential region h. Under the “strong sampling” (Tenenbaum and Griffiths, 2001) assumption that each event is independently sampled from the cause,

$$p(\{x_i\}_k|h') = \prod_{i: x_i \in \text{cause} k} p(x_i|h') \qquad \text{(Equation 4.4)}$$

Assuming uniform sampling from the consequential region, the likelihood that event x_i is sampled from consequential region h' is inversely proportional to the width of the region $|h'|$ if the event is within the consequential region, and zero otherwise:

$$p(x_i|h') = \begin{cases} \frac{1}{|h'|} & \text{if } x_i \in h' \\ 0 & \text{otherwise} \end{cases} \qquad \text{(Equation 4.5)}$$

This gives, for the likelihood of the current observation under latent cause k,

$$p(x_t \in k|\{x_i\}_k) = \frac{\sum\limits_{h':\{x_i\}_k, x_t \in h'} \frac{1}{|h'|^{n_k}} p(h')}{\sum\limits_{h:\{x_i\}_k, x_t \in h} \frac{1}{|h|^{n_k}} p(h)} \qquad \text{(Equation 4.6)}$$

The posterior probability of latent cause k is then updated using Bayes rule:

$$p(Z = k|x_t) = \frac{p(x_t|Z = k)p(Z = k)}{\sum\limits_{k'=1}^{K} p(x_t|Z = k')} \qquad \text{(Equation 4.7)}$$

Because the Bayesian inference process becomes intractable as the number of observations grows, we approximated the process using a particle filter (Gershman et al., 2010; Fearnhead, 2004) in which each particle maintains a single maximum *a posteriori* estimate of the assignment of observations to latent causes, rather than maintaining the full posterior distribution. We ran 8 simulations (four with $\alpha = 0.25$ and four with $alpha = 0.5$) using 50 particles each. As the number of true clusters was one for the training sequence and two for the test sequence, the concentration parameters were chosen such that the prior would produce one ($\alpha = 0.25$) or two ($alpha = 0.5$) clusters after 40 trials.

After observing the sequence of donations and inferring the donors' latent causes, the model estimated the average donation by taking the mean of the average donation of each latent cause weighted by the log-transformed number of donors that were assigned to the cause, instead of the true mean of all the experienced donations. We used the log-number of events rather than the true number to account for participants' losing precision over counts as more events are experienced. This makes causes with a smaller number of donors more impactful and causes with a larger number of events less impactful than linear weights.

4.2.2 Results and Discussion

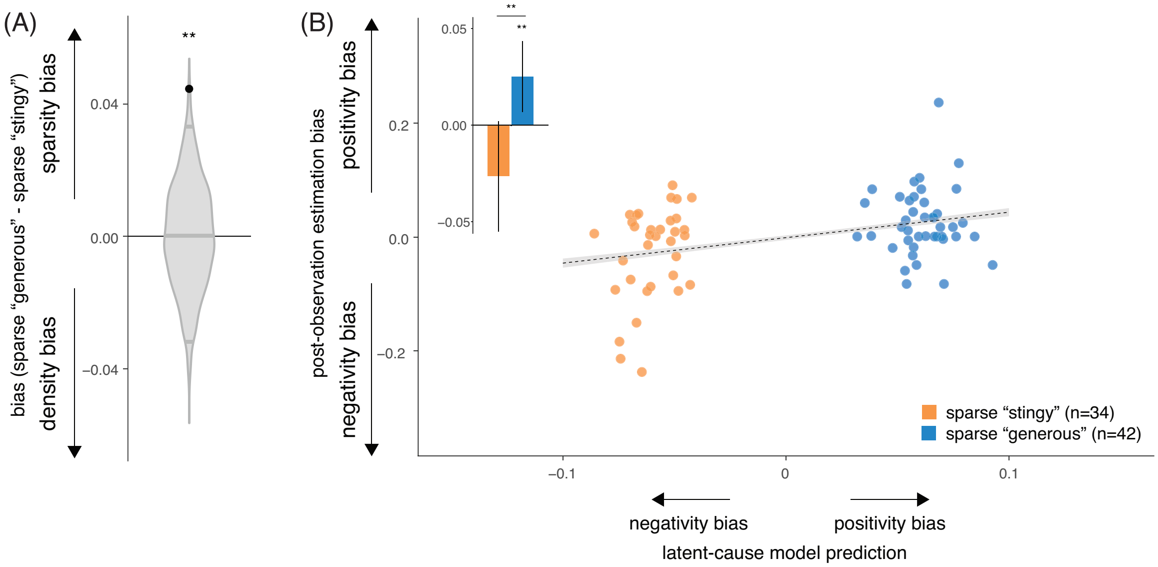

Figure 4.2: (A) The average estimate shows a sparsity bias. A permutation test in which condition labels were shuffled 2000 times (gray shaded) showed that the observed difference between the normalized post-observation estimate in the sparse “generous” and the “stingy” conditions in (black dot) was outside the null distribution confidence interval (marked with horizontal lines within the violin plot). (B) Comparison between simulation results and behavioral results. Each dot represents the estimation bias predicted by a latent-cause model that experienced the subject-specific sequence (x-axis) and the behavioral unnormalized post-observation estimation bias (i.e., post-estimation minus the true mean) of that subject (y-axis). Linear regression showed that estimate biases simulated by the latent-cause inference model predicted the estimate biases observed in the experiment (illustrated by the dashed line; see text). The inset shows the post-observation estimation bias difference between the sparse “stingy” (orange) and sparse “generous” (blue) conditions. Error bars indicate 95% confidence intervals. $^{**}p < 0.01$.

Effects of sparsity manipulation

We first investigated whether sparsity manipulates leads to biases in estimation. We normalized the estimates at each stage by the given scale, such that biases could range from -1 to 1. To account for individual differences in the prior beliefs, we normalized estimates by subtracting individual guesses prior to the test sequence observation (Guess 2) from the post-observation estimate. The normalized estimate was the main dependent variable of interest.

A two-tailed t-test on the normalized estimate difference between sparsity conditions showed a significant difference ($t(74) = -2.744$, $p = 0.008$, Cohen's $d = -0.633$, 95% CI = $[-0.077, -0.012]$), where the normalized estimate (i.e., the post-observation estimate minus "Guess 2") in the sparse "stingy" condition ($M = -0.020$) was significantly below the sparse "generous" condition ($M = 0.024$). A permutation test in which condition labels were shuffled 2000 times further showed that the observed estimate difference between sparsity conditions was outside the null distribution confidence interval (Figure 4.2A; estimated normalized difference = 0.044; 95% null distribution CI $[-0.030, 0.033]$, $p = 0.006$).

Post-observation biases

We then tested if post-observation estimates show a sparsity bias (i.e., a positivity bias in the sparse "generous" condition and vice versa) from the true mean in each sparsity condition (4.2B inset). One-sample t-tests showed that there was a significant sparsity bias in the sparse "generous" condition ($M = 0.025$; $t(41) = 2.772$, $p = 0.008$, Cohen's $d = 0.428$, 95% CI = $[0.007, 0.044]$), whereas we found no statistically significant effect of negativity bias in the sparse "stingy"condition ($M = -0.026$; $t(33) = -1.884$, $p = 0.068$, Cohen's $d = -0.323$, 95% CI = $[-0.055, 0.002]$). These post-observation estimate biases were significantly different across conditions ($t(74) = -3.196$, $p = 0.002$, Cohen's $d = -0.737$, 95% CI = $[-0.084, -0.019]$), where the sparse "stingy" condition was more negatively biased than the sparse "generous" condition.

Pre-observation biases

To further ensure that the condition difference at the post-observation estimate was attributable to the sequence manipulation, we tested for differences between conditions in the initial, pre-observation guesses (Guess 2). As expected, there was no evidence for a statistically significant difference in Guess 2 between conditions (sparse "stingy" condition $M = -0.006$; sparse "generous" condition $M = 0.001$; $t(74) = -0.912$, $p = 0.365$, Cohen's $d = -0.210$, 95% CI = $[-0.023, 0.009]$).

Comparison with inference model prediction

The model showed the sparsity bias (mean condition difference = 0.12, $t(74) = 43.622$, $p < 0.001$), as in the behavioral results. To take advantage of the property of Bayesian approximation that can capture an order effect, whereby presentation order influences groupings of the observed values (Austerweil and Griffiths, 2013), we provided the donations values to the model in the same order that each participant observed them. The model could thus make specific predictions about the estimation bias per each individual participant. We tested whether the individual behavioral biases would be predicted by the simulated biases. A linear regression analysis showed that the estimate biases simulated by the latent-cause inference model predicted the estimate biases behaviorally observed in the experiment ($\beta = 0.028$, $p < 0.001$, $\eta_p^2 = 0.145$; Figure 4.2B dashed line). This relationship was marginally significant even when controlling for the sparsity condition ($\beta = 0.080$, $p = 0.061$, $\eta_p^2 = 0.047$).

4.3 Experiment 1B

The core assumption of our latent-cause model is that the estimate of the average donation is made by averaging over latent causes' summary values. If participants knew in advance that they would only need to keep track of one summary value—the mean of all observations, they might update a running average rather than grouping donors into latent structures (Eyal et al., 2011). Therefore, in Experiment 1B (N = 22), we interrupted the latent-cause inference process by asking participants to report their average estimate on every trial (all other procedures were identical to Experiment 1A). We predicted that, by requiring participants to keep track of the overall mean, there would be no biases toward the sparse area.

4.3.1 Methods

Participants

Twenty-two participants (10 in the sparse "stingy" condition; 12 in the sparse "generous" condition) were recruited using Amazon Mechanical Turk (MTurk). Exclusion criteria were identical to Experiment 1A.

Task and Procedure

We added an average estimation task upon each observation. After observing and logging each donation, participants were asked to estimate the average thus far. All other procedures and materials were identical to those used in Experiment 1A.

4.3.2 Results and Discussion

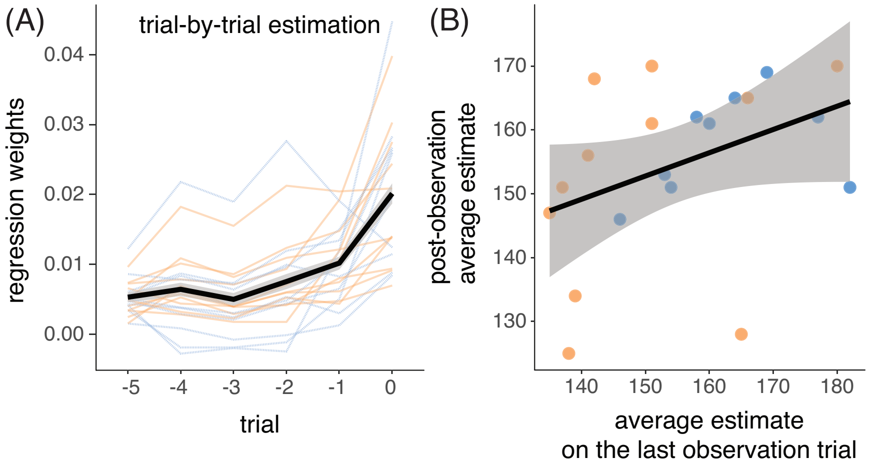

Figure 4.3: A) Recency biases in trial-by-trial estimates. Linear regression showed a recency bias, where more recent observations contributed more strongly to the current estimate (Orange lines: sparse “stingy” condition; blue lines: sparse “generous” condition). (B) Recency bias and final post-observation average estimate. The trial-by-trial estimate on the last trial (x-axis) was marginally predictive of the post-observation (and post filler-task) estimate (y-axis). Each dot represents one subject, grey shading: 95% confidence intervals.

Effects of sparsity manipulation

When subjects were required to estimate the mean donation after every donation, the normalized estimates were no longer significantly different across conditions (sparse “stingy” $N = 10$, $M = 0.032$; sparse “generous” $N = 12$, $M = 0.012$; $t(20) = 0.984$, $p = 0.337$, Cohen’s $d = 0.421$, 95% CI = $[-0.023, 0.063]$), with the numerical difference in the direction opposite to the latent-cause model prediction. We tested if these differences between the sparsity conditions interact with the tracking manipulation, pooling data from Experiment 1A (end-of-sequence estimation) and Experiment 1B (trial-by-trial estimation). A two-way ANOVA showed a significant interaction between the sparsity and tracking conditions ($F(1, 94) = 4.056$, $p = 0.047$, $\eta_p^2 = 0.04$), suggesting that the sparsity in the distribution induces biases only when the average values are not tracked on a trial-by-trial basis.

The relationship between trial-by-trial estimates and post-observation estimates

Further, we explored the relationship between observation, trial-by-trial estimates, and final post-observation estimates in Experiment 1B. Linear regression showed that more recent observations contributed more strongly to the current estimate (Figure 4.3A). The trial-by-trial estimate on the last trial was also marginally predictive of the post-observation average estimate (Figure 4.3B; $\beta = 5.362$, $p = 0.082$, $\eta_p^2 = 0.158$). Together, the results suggest that the overall average estimate is derived via a different strategy when there is a clear goal of tracking the average value, supporting our hypothesis that latent-cause inference could be the mechanism by which sparse events become overweighted in the overall estimate.

4.4 Experiment 2A

In Experiment 1, the sparsity manipulations induced biases in average estimation when there is no explicit goal of tracking the average value. However, there are two alternative explanations for the bias we observed. First, even if subjects did not infer multiple causes for the sparse events but rather perfectly inferred that there are two latent causes (“stingy” and “generous”), log-weighted average of the mean values of these two causes would have resulted in an estimate that is biased toward the cause that has fewer events. That is, the log-weighted means of the “stingy” and “generous” causes would show biases that are in line with the latent-cause inference model prediction, as the fewer events in the sparse latent-cause gain more prominence. To address this, in Experiment 2, we equated the frequency of events that are generated by the “stingy” and “generous” causes and only manipulated the variance of the event distributions (Figure 4.4A).

Second, Pearce and Hall (1980) suggest that more surprising events will update values of an entity to a greater degree. Because events in the sparse area elicits higher prediction errors (surprises), they may have a greater impact on the learned averages estimate (formally, these surprising events will have a higher learning rate). To adjudicate between the latent-cause inference model and the Pearce-Hall dynamic-learning-rate model, in Experiment 2, we chose a specific presentation order where the distributions from which observed values are drawn quasi-alternated between the dense and the sparse, and the end of sequence was predominantly populated with values from the dense distribution (Figure 4.4A). Alternating between the dense and the sparse distributions made the trials from both distributions similarly surprising on average, eliciting similar levels of prediction error and therefore equating attention to both distributions in the Pearce-Hall model. In addition, as values from the dense distribution were observed just prior to average estimation, an error-driven learning model would show a density bias due to the enhanced effect of recent experiences in such models (Figure 4.5). Given these properties of the chosen presentation order, the Pearce-Hall model with dynamic learning rates predicted a density bias, while latent-cause inference still predicted a sparsity bias (Figure 4.4B).

To this end, Experiment 2A used fixed observation sequences that equated the frequency of “stingy” and “generous” events, under a cover story that each observation was how much a community member was willing to pay for a charity event.

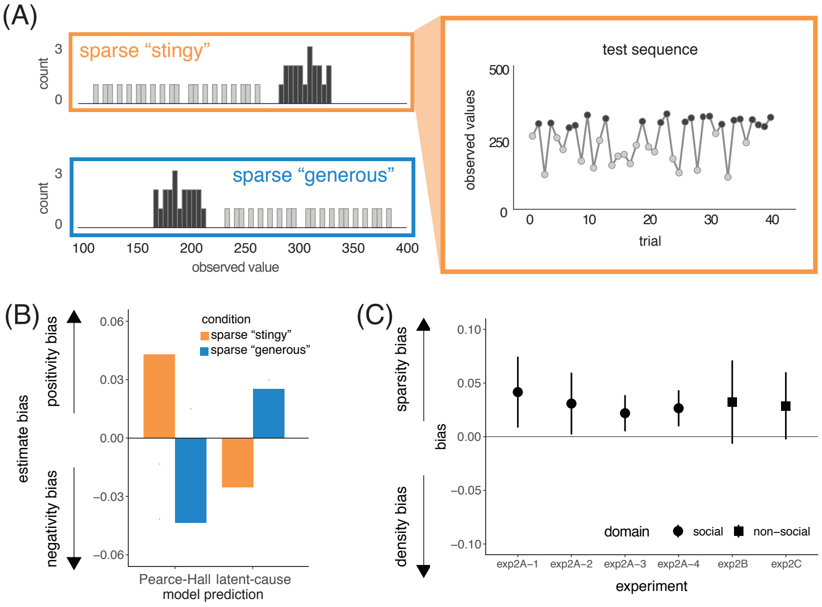

Figure 4.4: Sparsity biases when rarity was matched between the sparse and dense causes. (A) Distributions and test sequence. In Experiment 2, the frequency of "generous" (dark gray) and "stingy" (light gray) donations was matched (Left panel). For each condition, we also chose a specific sequence such that "generous" and "stingy" donations would elicit similar levels of prediction error (Right panel). In this sequence, as the two distributions mostly alternated throughout the sequence, making both donation amounts locally surprising. Finally, the sequence ended with values drawn from the dense distribution (see text). (B) Model predictions and empirical results. For the sequences in Experiment 2, the Pearce-Hall model (left) and latent-cause model (right) predicted biases in opposite directions. (C) The empirical results were in line with the latent-cause model prediction, showing a sparsity bias. Error bars indicate 95% confidence intervals.

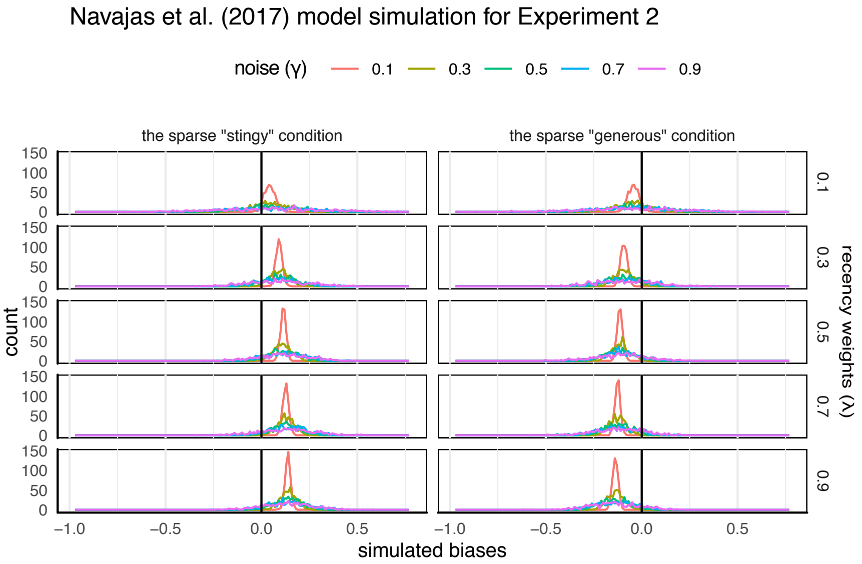

Figure 4.5: Alternative model simulation results for Experiment 2. In addition to a learning model that scales learning rates by surprise Pearce and Hall (1980), we simulated a model that takes into account the observation's variance, scaling decision noise γ by the variance of the observation sequence (Navajas et al., 2017), using the stimulus sequences from Experiment 2A. The stimulus sequences were designed such that the values from the dense distribution were observed just prior to average estimation, thereby leading a recency-weighted model to show a density bias due to the enhanced effect of recent experiences, while leaving a latent-cause inference model to show a sparsity bias. The simulation results showed a density bias in both the sparse "stingy" (left) and "generous" (right) conditions for different levels of recency-weights λ and decision noises γ and were opposite to our empirical results that showed a sparsity bias.

4.4.1 Methods

Participants

The total of 626 participants were included in Experiment 2A (2A-1 $N = 70$, sparse "stingy" $N = 26$, sparse "generous" $N = 44$; 2A-2 $N = 67$, sparse "stingy" $N = 28$, sparse "generous" $N = 39$; 2A-3 $N = 260$, sparse "stingy" $N = 133$, sparse "generous" $N = 127$; 2A-4 $N = 229$, sparse "stingy" $N = 118$, sparse "generous" $N = 127$). For Experiment 2A-3, sample size was chosen from power analysis based on Experiment 2A-1 and 2A-2. For Experiment 2A-4, we took a Bayesian approach and collected minimum of 50 usable participants in each condition and continued data collection until we reached one of three criteria : (1) we obtain the Bayes Factor of 10 in favor of H1 (normalized estimate in sparse “stingy” condition < normalized estimate in sparse “generous” condition) and against H0 (no difference in normalized estimates between sparsity conditions), (2) we obtain the Bayes Factor of 10 in favor of H0 and against H1, or (3) we reached the maximum number of participants (500 usable participants in each condition). This procedure was pre-registered.

Exclusion criteria were identical to Experiment 1A and Experiment 1B, except that subjects who missed the 5-second response window for logging the amount were excluded.

Task and Procedure

We used a different cover story to generalize our results. In Experiment 2A, participants were told that they were selling coffee for charity events at community fairs in different towns (“Lambtonville” and “Brookfield”) and taking coffee orders. The customers could pay in tokens as they wish, and the participants’ task was to log the payment amount for each customer. The customer names were shown in the prompts (e.g., ”Brennan: 218 tokens for Cappuccino”). We instructed participants to pay attention to both the names and the payment amount, as some pairs of name and payment amount will be tested at the end. This was to orient them to pay attention to the task (For Experiments 2A-1 and 2A-2, we asked participants to report the payment amount for given customers at the end of the experiment. For Experiment 2A-3 and 2A-4, we did not test participants’ memory. In all cases, we did not analyze these data, as they were outside the scope of our interest).

Tokens did not appear on the screen as visual cues, and the response was made either by moving a slider ranging from 1 to 500 tokens (Experiment 2A-1, 2A-2; to help participants make a response within the response window, the slider snapped to the correct number when the distance between the marker and the target was less than 5 tokens) or by typing in the number (Experiment 2A-3, 2A-4), with a 5-second time limit. Participants got a 50 cents bonus if they did not miss any orders.

Critically, the sparsity of "stingy" and "generous" was manipulated by variance alone. In both conditions, the number of customers generated from stingy and generous causes were matched to 20. In the sparse stingy customer condition, the "stingy" customers' payment amounts were more variable than the "generous" donors (20 "stingy" donors $M = 188$, $SD = 47.29$; 20 "generous" donors $M = 308$, $SD = 13.73$), and vice versa in the sparse generous customer condition (20 "generous" donors $M = 312$, $SD = 47.29$; 20 "stingy" donors $M = 192$, $SD = 13.73$).

The order of payment values was chosen such that the latent-cause inference model and the Pearce-Hall model predict the opposite biases.

Pearce-Hall model

For Pearce-Hall model simulations, donation amounts were normalized to the maximum potential amount (i.e., the maximum amount on the response slider bar) and then introduced to the Pearce-Hall model (Pearce and Hall, 1980). The value estimate v was updated according to:

$$v_{t+1} = v_t + \alpha_{t+1} \times S \times x_t \qquad \text{(Equation 4.8)}$$

where α is the associability parameter, S denotes salience of the cue, and x represents the observed amount. The key component of the Pearce-Hall model is that the associability α is updated according to the absolute prediction error (the difference between the observed and expected values), with a learning rate η:

$$\alpha_{t+1} = (1 - \eta) \times \alpha_t + \eta \times |x_t - v_t| \qquad \text{(Equation 4.9)}$$

This means that more surprising events have greater impact on the overall value estimates. We ran simulations with salience parameter S ranging from 0.1 to 1, and learning rate η ranging from 0.1 to 1. The evaluation of a group v was made using each combination of the two parameters.

4.4.2 Results and Discussion

Effects of sparsity manipulation

In Experiment 2A-1 ($N = 70$), a two-sample t-test showed that the normalized estimate in the sparse “stingy” condition ($M = -0.023$, $N = 26$) was significantly lower than in the sparse “generous” condition ($M = 0.018$, $N = 44$; $t(68) = -2.511$, $p = 0.014$, Cohen’s $d = -0.621$, 95% CI = $[-0.075, -0.009]$). Further, a permutation test in which condition labels were shuffled 2000 times further supported that the observed sparsity bias effect was above the null distribution confidence interval (normalized condition difference = 0.044; 95% null distribution CI $[-0.035, 0.033]$, $p < 0.05$). Critically, these results were in the direction that the latent-cause inference model predicted and opposite to the predictions of the Pearce-Hall model.

To strengthen this finding, we ran three independent sets of replications (Experiment 2A-2 $N = 67$; 2A-3 $N = 260$; 2A-4 $N = 229$, pre-registered). All three experiments replicated the main finding that the sparse “stingy” condition’s normalize estimates (Experiment 2A-2 $N = 28$, $M = -0.014$; 2A-3 $N = 133$, $M = -0.009$; 2A-4 $N = 118$, $M = -0.005$) were below the sparse “generous” condition’s normalize estimates (Experiment 2A-2 $N = 39$, $M = 0.017$; 2A-3 $N = 127$, $M = 0.013$; 2A-4 $N = 111$, $M = 0.021$). The difference between conditions were significant (Experiment 2A-2 $t(65) = -2.137$, $p = 0.036$, Cohen’s $d = -0.529$, 95% CI = $[-0.060, -0.002]$; 2A-3 $t(258) = -2.556$, $p = 0.011$, Cohen’s $d = -0.317$, 95% CI = $[-0.039, -0.005]$; 2A-4 $t(198.54) = -3.098$, $p = 0.002$, Cohen’s $d = -0.407$, 95% CI = $[-0.043, -0.010]$). An additional meta-analysis across Experiment 2A (a, b, c,

d) using Bayes Factor showed BF10 of 17732.306, indicating the data being 17732.3 times more likely under the hypothesis that the normalized estimates are different across sparsity manipulations than the null hypothesis.

Post-observation biases

We then tested whether post-observation estimates in each condition show a sparsity bias from the true mean across these four replication data sets, using meta-analytic Bayes Factor analyses on the two-tailed t-tests (Figure 4.6). Negativity biases observed in the sparse "stingy" condition (Experiment 2A-1 $M = -0.027$; 2A-2 $M = -0.023$; 2A-3 $M = -0.012$; 2A-4 $M = -0.017$) were 636.6 times more likely under the hypothesis there is a bias than a null hypothesis ($BF10 = 636.584$), providing extremely strong evidence for a bias. In the sparse "generous" condition, positivity biases (Experiment 2A-1 $M = 0.006$; 2A-2 $M = 0.008$; 2A-3 $M = 0.006$; 2A-4 $M = 0.014$) were 4.6 times more likely under the hypothesis there is a bias than a null hypothesis ($BF10 = 4.556$), providing moderate evidence for a bias.

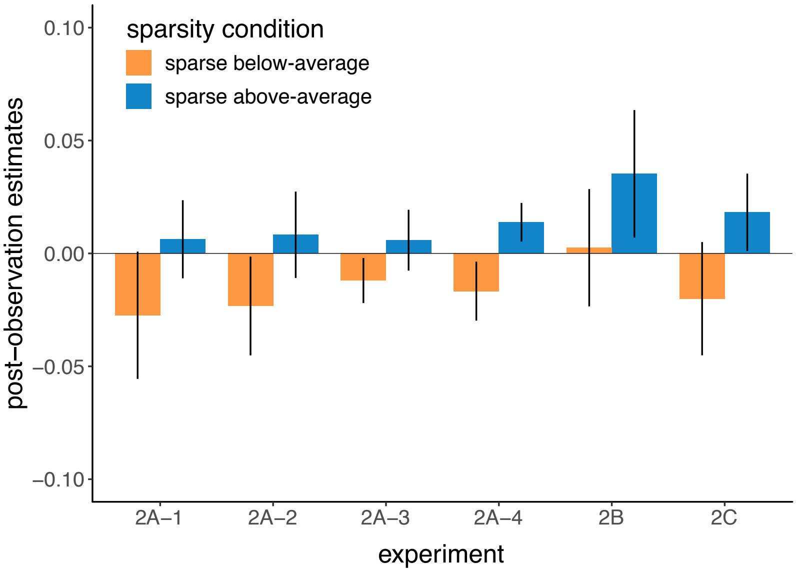

Figure 4.6: Negativity biases in the sparse "stingy/below-average" condition was marginally significant (Experiment 2A-1 $M = -0.027$, $t(25) = -2.001$, $p = 0.056$, Cohen's $d = -0.392$, 95% CI $= [-0.056, 0.001]$) or significant (Experiment 2A-2 $M = -0.023$, $t(27) = -2.185$, $p = 0.038$, Cohen's $d = -0.413$, 95% CI $=$ $[-0.045, -0.001]$; Experiment 2A-3 $M = -0.012$, $t(132) = -2.382$, $p = 0.019$, Cohen's $d = -0.207$, 95% CI $= [-0.022, -0.002]$; Experiment 2A-4 $M = -0.017$, $t(117) = -2.532$, $p = 0.013$, Cohen's $d = -0.233$, 95% CI $= [-0.030, -0.004]$) in the social domain. However, such biases was not significantin the non-social domain (Experiment 2B $M = 0.003$, $t(37) = 0.197$, $p = 0.845$, Cohen's $d = 0.032$, 95% CI $=$ $[-0.023, 0.029]$; Experiment 2C $M = -0.020$; $t(50) = -1.605$, $p = 0.115$, Cohen's $d = -0.225$, 95% CI $= [-0.045, 0.005]$). Positivity biases in the sparse "generous" condition was not significant in three replications (Experiment 2A-1 $M = 0.006$, $t(43) = 0.731$, $p = 0.469$, Cohen's $d = 0.110$, 95% CI $= [-0.011, 0.024]$; Experiment 2A-2 $M = 0.008$, $t(38) = 0.874$, $p = 0.388$, Cohen's $d = 0.140$, 95% CI $=$ $[-0.011, 0.027]$; Experiment 2A-3 $M = 0.006$, $t(126) = 0.860$, $p = 0.391$, Cohen's $d = 0.076$, 95% CI $= [-0.008, 0.019]$) while significant in one (Experiment 2A-4 $M = 0.014$, $t(110) = 3.196$, $p = 0.002$, Cohen's $d = 0.303$, 95% CI $= [0.005, 0.022]$). In the non-social domain, positivity bias was significant in both Experiment 2B ($M = 0.035$, $t(42) = 2.519$, $p = 0.016$, Cohen's $d = 0.384$, 95% CI $= [0.007, 0.063]$) and Experiment 2C ($M = 0.018$, $t(49) = 2.134$, $p = 0.038$, Cohen's $d = 0.302$, 95% CI $= [0.001, 0.035]$).

4.5 Experiments 2B, 2C

We further sought to investigate whether the sparsity bias in the social domain emerges from a domain-general inference process. If the process by which overall estimation is made is different across domains, we would expect to observe an interaction in the bias between the sparsity conditions and the domains (social and non-social). Specifically, in the non-social domain, we would observe no difference between the sparsity conditions or a difference in the opposite direction (i.e., the density bias), following the Pearce-Hall model's prediction. However, if the bias arises from a fundamental inference process, the sparsity in the observed values should lead to biases regardless of domains. To this end, Experiment 2B and Experiment 2C used non-social cover stories (the weight of coffee beans that customers buy from different towns in Experiment 2B; the amount of slot machine earnings from different casinos in Experiment 2C). All other procedures were identical to their social counterpart (Experiment 2A), and the sparse "below-average" and "above-average" conditions here respectively corresponded to the sparse "stingy" and "generous" conditions in Experiment 2A.

4.5.1 Methods

Participants

The total of 182 subjects participated in Experiment 2B ($N = 81$, sparse "stingy" $N = 38$, sparse "generous" $N = 43$) and Experiment 2C ($N = 101$, sparse "stingy" $N = 51$, sparse "generous" $N = 50$). Exclusion criteria were identical to Experiment 2A.

Task and Procedure

To investigate the sparsity bias in the non-social domain, we changed the cover story such that participants were logging weights of coffee beans for customers in supermarkets in different towns (Experiment 2B) or logging slot machine earnings in different casinos (Experiment 2C; participants' compensation did not depend on observed earnings to avoid those amounts from playing the role of personally-relevant rewards). Critically, the stimuli sequences were the same as Experiment 2A, where the sparsity was manipulated by variance. All procedure was identical to Experiment2A-3 and d, where responses were made in a text box.

4.5.2 Results

Effects of sparsity manipulation

Experiment 2B ($N = 81$) and Experiment 2C ($N = 101$) showed that the sparse "below-average" condition's normalized estimates (Experiment 2B $N = 38$, $M = 0.015$; Experiment 2C $N = 51$, $M = 0.001$) were numerically below the sparse "above-average" condition (Experiment 2B $N = 43$, $M = 0.048$; Experiment 2C $N = 50$, $M = 0.030$), although they did not reach statistical significance from two-tailed two-sample t-tests (Experiment 2B $t(79) = -1.654$, $p = 0.102$, Cohen's $d = -0.368$, 95% CI $= [-0.071, 0.007]$; Experiment 2C $t(99) = -1.822$, $p = 0.071$, Cohen's $d = -0.363$, 95% CI $= [-0.060, 0.003]$).

Interaction between domains and sparsity manipulation

To test whether the conditional difference in the overall estimate interacted with the domain, we ran a mixed-effects linear regression model predicting normalized estimates with the sparsity and social ("social" or "non-social") conditions as fixed effects and experiments (Experiments 2A-1, 2A-2, 2A-3, 2A-4, 2B, 2C) as random effects, pooling data across Experiment 2 (N = 808). Tests of significance using Satterthwaite's approximation showed no significant interaction between the sparsity and social conditions (β = 0.005, SE = 0.012, $t(804)$ = 0.393, p = 0.695). Bayes Factor analysis provided moderate evidence that there is no interaction between event domain (social vs non-social) and the sparsity conditions ($BF01 = 7.637$).

Post-observation biases

We examined each condition's sparsity bias in the non-social domain (Figure 4.6). While there was a strong evidence for a positivity bias in the sparse "above-average" condition ($BF10 = 16.153$; null hypothesis cohen's d = 0), there was a moderate evidence that there is no bias in the sparse "below-average" condition ($BF01 =$ 4.902).

Pre-observation biases

The lack of negativity bias in the sparse "below-average" condition in the non-social domain may be due to prior expectation. That is, if prior beliefs about the events are skewed towards smaller values under the non-social domain, positive events with larger values may weigh more heavily in overall estimation of observed events. To this end, we ran a mixed-effects linear regression model predicting Guess 2 with the social conditions as fixed effects and experiments as random effects. A test of significance using Satterthwaite's approximation showed that prior beliefs in the non-social domain ($M = -0.015$) was significantly below priors in the social domain ($M = -0.007$; $\beta = 0.007$, $SE = 0.003$, $t(7.308) = 2.539$, $p = 0.037$).

4.6 Experiment 3

A potential mechanism for a general negativity bias is that each observation is perceived on a logarithmic rather than linear scale. In this case, small numbers would have greater relative significance than larger numbers, leading the overall estimation to be biased to below the true (linear) mean. This would predict negativity biases regardless of sparsity manipulations.

To test the logarithmic hypothesis directly, we ran a control experiment matching the two conditions (sparse “stingy” and “generous”) for the arithmetic mean of log-transformed donation values.

4.6.1 Methods

Participants

Thirty-nine subjects participated in Experiment 3 (sparse “stingy” $N = 19$, sparse “generous” $N = 20$). Exclusion criteria were identical to Experiment 1A.

Task and Procedure

Here, we matched the two conditions (sparse “stingy” and “generous”) for the arithmetic mean of log-transformed donation values at 100 (Figure 4.7 solid lines). The means of log-transformed values were both lower than the true linear-scale mean (Figure 4.7 dotted lines; the sparse “stingy” condition mean = 108.65; the sparse “generous” condition mean = 111.65). Other procedures were identical to Experiment 1A, and the values were normalized by the scale (300) for the following analyses.

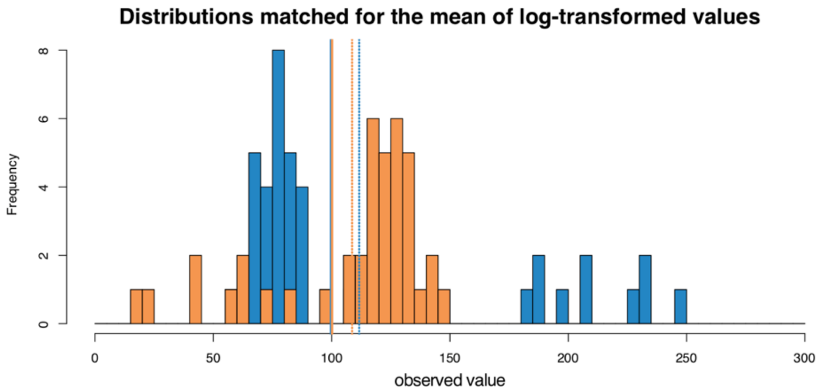

Figure 4.7: In Experiment 3, the distributions for sparse "stingy" (orange) and sparse "generous" (blue) conditions were determined such that the arithmetic means of *log-transformed observed values* (solid lines) were matched (rather than matching the arithmetic means of the observed values themselves, as in all our other experiments). The true means of linear-scale values (dotted lines) were higher than the means of the log-transformed values in both conditions.

4.6.2 Results and Discussion

To test if estimation biases are driven by the logarithmic scale, we first compared the post-observation estimates to the mean of the log-transformed values. The empirical estimates were significantly different from the mean of the log-transformed values (mean difference = 0.060, $t(38) = 4.543$, $p < 0.001$, Cohen's $d = 0.727$, 95% CI = $[0.033, 0.086]$). Specifically, the estimates were higher than the mean of the log-transformed values in both the sparse "stingy" condition ($N = 19$; mean difference = 0.034, $t(18) = 2.203$, $p = 0.041$, Cohen's $d = 0.505$, 95% CI = $[0.002, 0.067]$) and the sparse "generous" condition ($N = 20$; mean difference = 0.084; $t(19) = 4.232$, $p < 0.001$, Cohen's $d = 0.946$, 95% CI = $[0.042, 0.126]$).

A key prediction of the logarithmic hypothesis is that the estimated mean would be *lower* than the true linear-scale mean, regardless of the sparsity manipulation. Both our latent cause model and the logarithmic hypothesis predict a negative bias for the sparse "stingy" condition. However, if estimation biases emerge as a result of latent-cause inference, the sparse "generous" condition should show a positivity bias that the logarithmic hypothesis cannot predict. A two-tailed t-test showed that the post-observation estimate was positively biased in the sparse "generous" condition (mean bias = 0.044, $t(19) = 2.210$, $p = 0.040$, Cohen's $d = 0.494$, 95% CI = $[0.002, 0.085]$), supporting the latent cause inference hypothesis.

4.7 General Discussion

Together, these experiments demonstrated that overall estimation of a quantity is biased toward the value of events that are rare and/or more variable. Comparing human participants' biases to simulated estimation biases from a semi-rational latent-cause inference model suggested that the behavioral results can be attributed to a process of inferring latent causes for observations and estimating the overall average by averaging over these causes.

Our results are in line with empirical findings in social cognition research showing that, given rarity (Fiske, 1980) and variability (Unkelbach et al., 2008; Alves et al., 2017b) of negative events, negative information can have a higher impact in impression formation and updating (Rozin and Royzman, 2001). To test our hypothesis that the distributional sparsity is driving such bias, we used donation events in a positive monetary domain and manipulated sparsity of below-average (relatively negative) or above-average (relatively positive) events. This was a strong test of our hypothesis, as we avoided events in the negative domain altogether so as to not confound our findings with subjective value differences for monetary wins and losses (i.e., the fact that losses loom larger than gains of the same amount17). We expect that the effects we observed would occur even more strongly when the valence of the stimuli varies across the full spectrum of negativity and positivity, since there are features of negative events that make biased processing of negative stimuli adaptive. For instance, an untrustworthy person can impose a risk on our well-being, and thus it would be wise to avoid such risk. If we choose to avoid a presumably bad person, we lose opportunities to update our impression and effectively make our samples of that person's behaviors sparse, leading to a negativity bias (Denrell, 2005). On the other hand, we may update our impressions differently when we observe bad behaviors, leaving more chance to forgive potentially bad targets (Siegel et al., 2018). Moreover, valence may not change monotonically with magnitude for some behaviors; for example, talking too much could be as negative as talking too little, while donating more money is generally positive (Alves et al., 2017b). The different weighting of gain and loss is also a potential source of asymmetry between the negative and positive-valenced domains. Prospect theory suggests that the value function is steeper for losses than for gains, meaning that a loss of $5 looms much larger than a gain of the same amount (Kahneman and Tversky, 1979). Future studies can explore potential interaction between sparsity and the valence of events in social and non-social domains.

We explored whether the inference processes that give rise to the sparsity bias are domain general or uniquely social, by using various social and non-social scenarios. There was moderate evidence that the sparsity bias was not different across domains, suggesting domain-general inference processes. However, our social and non-social cover stories may diverge in other important ways, especially with regard to prior beliefs. Although we presented participants with a training sequence to neutralize their prior beliefs, the training may not be enough to adjust people's expectations for various situations, as these were built through a lifetime of experiences. This may be the reason that there was a difference in "Guess 2" between the social and non-social domains. In a donation scenario, the donation amounts observed during the first training sequence may be a more informative ground for processing the next set of observations, as they form a social norm for generosity that the members of the community should follow. This would make the prior beliefs sharper, and any donations outside the normal range could be perceived as good or bad. On the other hand, our non-social scenarios (how many grams of coffee beans are purchased, or how much slot machines return) may have evoked priors whose values are lower than the donation scenario, and thus every win in a casino or sale of coffee beans, even if below average, may have continued to be seen as a positive event, leading to a stronger positivity bias. This would be a potential explanation for the absence of a negativity bias when below-average events were sparse in the non-social scenarios. Future research should investigate how prior beliefs differ across situations, and whether priors under social situations have unique characteristics that alter the inference process or allow it more flexibility. Quantifying individuals' prior beliefs can also allow the model to predict individual differences in negativity biases, which indeed tend to be stable within an individual across time (Ito and Cacioppo, 2010). For instance, an individual who has experienced largely negative life experiences may have a prior expectation that positive events are sparse, and therefore will weigh positive events more heavily in their overall estimation.

Our study also differs from previous impression-formation studies in that we assessed the impression of a group rather than an individual. When forming impressions about an individual, we tend to assume more unity and coherence in their behaviors than we do about groups, drawing inferences about dispositional properties (Hamilton and Sherman, 1996; Asch, 1946; Jones and Davis, 1965). Group impression formation may rely on a different process from person impression formation (Fiske and Neuberg, 1990) , which involves representing multiple individual experiences, or exemplars (Smith and Zárate, 1992), depending on perceived entitativity (i.e., the degree of having the properties of an entity; Campbell, 1958; Lickel et al., 2000) of a group. The level of judgment, thus, can be intermediate between lumping all individuals into one category and representing each individual as its own entity. This resembles the tension between only representing prototypes of a category (prototype model Reed, 1972) and preserving all exemplars (exemplar model Nosofsky, 1986; for a review, see Hilton and von Hippel, 1996). The latent-cause inference model can be seen as an intermediate between the two alternatives, achieving either end of the spectrum by varying a single parameter that governs creating a new cluster (Sanborn et al., 2010), and thus could be a good model for group impression formation. That is, the model could provide a framework to further explore how entitativity influences group impression formation. For example, In Experiment 2B and Experiment 2C where the causal link between observed behavior and the groups was weak, the sparsity bias predicted by the latent-cause inference model was less pronounced. A town might not be a coherent entity to predict coffee sales as much as it is for predicting social norms such as generosity. That is, when evaluating a heterogeneous social group that we assume has a common latent causal property (such as intention) that generates individual observations, sparse experiences with the group can drive our overall impression of the group. In any case, if these biases are the results of fundamental inferential processes that partition our experiences into meaningful causal chunks, the model should hold true in individual impression formation as well. Given that there is a closer and more immediate causal link between an individual and their actions than between a group and the group members' actions, the sparsity bias effect would potentially even be stronger for individual impressions.

Experiment 1 suggests a way to reduce the sparsity bias. We showed that being required to think about the overall mean in every step of observation promotes unbiased estimationm, as evidenced by the interaction between the task requirements and the sparsity conditions. This suggests that we may be less affected by rare and variable interactions if we try to track a particular quality of another person's every time we interact with them, rather than leaving the judgment until later. This could be desirable in a situation where we want an unbiased evaluation, for instance during a hiring process. Nevertheless, placing people on a positive-negative scale is usually not the only goal in our rich day-to-day interactions, and we often need more flexible representation of our social counterparts.

Our model estimates the overall average by taking the log-weighted mean of latent causes' mean values. That is, the model assumes that the low frequency is overestimated and high frequency is underestimated when the overall mean is estimated based on the latent causes' mean donation, as each cause is weighted by the log-transformed number of donors that were assigned to the cause. This loss of precision is based on numerous studies showing that low frequency and/or probability is overestimated while high frequency and/or probability is underestimated (Zhang and Maloney, 2012). Although this type of precision loss assumed by our model is repeatedly found in the literature (Merten and Nieder, 2008; Dehaene et al., 1998), it is worth noting that the degree of sparsity bias depends on the degree to which the latent causes' frequency information is distorted. At the extreme end of the spectrum where latent cause frequency is perfectly kept, there will be neither a sparsity nor a density bias. At the other end of the extreme where the frequency information is completely lost such that a latent cause that generates only one observation has the same impact on the overall mean as a cause that generates the rest of the observations, there will be a much stronger sparsity bias. This could also be why we observed a stronger sparsity bias in the social scenarios such as donations than non-social situations, as we represent the groupings of people by relying on existing schema we already have from previous experiences with other people, thereby further losing precisions on frequency of encounters in this particular setting.

Another possibility is that the frequency is distorted in the inference process as well as in the averaging of inferred causes. The CRP prior is a rich-gets-richer process where a new event will be more likely to be assigned to a cause with a larger number of events already assigned to it than to unpopular causes, which requires counting how many events already belong to the cause. The aforementioned frequency distortion can occur in this counting step. This does not change the direction of the bias although the degree of the bias decreases. Additionally, similar distortions can also occur even when there is no inference involved. That is, if the frequency of the same value gets lost, the dense distribution may contribute less to the overall mean, due to the higher chance of showing the same value multiple times than the sparse distribution. However, this type of frequency distortion cannot account for the biases seen in Experiment 2A, as the exact values of donations were unique per trial, and there was no repetition. Furthermore, in Experiment 1A where the same outcome was repeated, if we counted only the first presentation, losing all following presentations, the mean of unique donations in the sparse "stingy" condition would be above the true mean and vice versa, which is the opposite direction to the sparsity bias observed in the data.

Finally, logarithmic transformation can occur in representing the donation amount as well. If the donation amount is perceived in a logarithmic scale, the final average estimation would show negativity bias. Experiment 3 matching the log-transformed mean, not the linear-scale mean of the "sparse stingy" and "sparse generous" conditions showed that the estimation biases was positive in both conditions, which was the opposite of the log-scale representation prediction.

Given the possible hypotheses about the precision loss when accounting for the number of observation in each latent cause, the ideal approach would be to fit these models to empirical data and compare which type of precision loss predicts our results the best. This is not possible in the current study as we collect only one estimate per participant. We chose the current design to prevent continuous reporting from altering the cognitive processes by which the latent causes are inferred and the final evaluation is made. Of course, this choice came at the expense of model-fitting; future work can characterize the online inference process, for instance, by probing groupings on a trial-by-trial basis.

In conclusion, we have shown that sparse experiences are segmented into a larger number of latent causes, which in turn bias the overall impression such that the rarer and/or more variable experiences are overweighted. Here, we showed the sparsity bias in a mildly social domain. We would expect that, in more realistic social scenarios that involve evaluations of others with real stakes at hand, the biases may manifest even more strongly. This cognitive bias could indeed be the core mechanism underlying the negativity bias in social evaluations.

Chapter 5

Conclusion

5.1 Summary

Latent causes are a normative way of organizing experiences such that they can be retrieved and generalized where relevant. In this thesis, I first demonstrated that memory retrieval is facilitated when the study and test environment share the structure that generates perceptual and cognitive features. However, experimentally manipulating such generative structures may not always lead subjects to infer distinct generative structures, holding separate internal representations. To address how latent causes structure memory, I offered a theoretical account for episodic memory organization that relies on inference of latent causes. To characterize the effects of clustering experiences through latent cause inference in decision-making, I used a Bayesian inference model to predict empirical biases in evaluations in scenarios where clusters/groups of observations are ambiguous, such as social evaluations.

5.2 Ongoing Work: quantifying latent-cause inference

The Bayesian Latent-cause inference model discussed in this thesis offers a framework that can quantify the process by which we group our experiences for future use. Quantification of the inference process could provide insights into individual differences in how people generalize from their past experiences to new ones. Specifically, parameters corresponding to each component of the inference process can explain different aspects of tendencies to over- or under- generalize across experiences. To this end, I have developed a novel task that probes clustering trial-by-trial, allowing for fitting parameters of the latent-cause inference model.

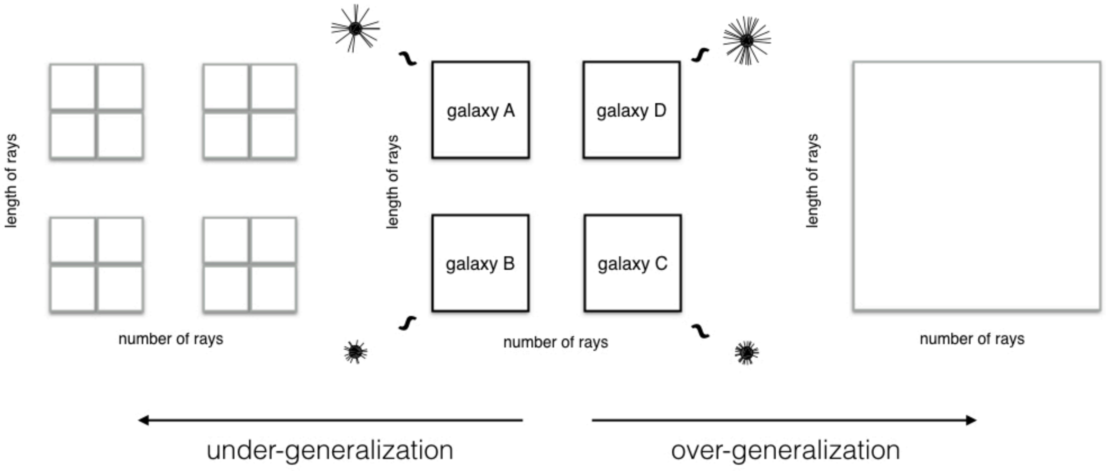

Figure 5.1: Four galaxies generate different types of stars with varying lengths and numbers of rays (middle panel). Subjects are asked to assign stars to an unknown number of galaxies, based on their perceptual features. Subjects may under-generalize, parsing the perceptual space more finely, assuming more galaxies than the ground truth (left). On the other hand, some individuals may over-generalize, failing to differentiate the galaxies (right).

In the inference task, subjects are told that they are classifying stars from different galaxies (latent causes). They are told that stars from a galaxy look similar to each other and have similar numbers and lengths of rays coming out of them. Their task is to classify the stars based on these perceptual features. Critically, to investigate individual differences in generalization, subjects are not told *a priori* how many galaxies they will encounter, and thus need to flexibly infer a new galaxy when they think that a star does not belong to any of the previous galaxies. This allows us to measure the probability that a new cluster is inferred, which corresponds to the concentration parameter α in the prior over clusters.

Importantly, stars from the same galaxy also tend to be presented in temporal proximity. Subjects may utilize these temporal statistics, decaying the influence of past stars on inference of the galaxy of the current star as a function of time. As a result, when seeing a star from a previous galaxy, subjects who have a high decay rate λ might fail to recognize the old galaxy, identifying a new galaxy instead. The decay parameter λ therefore explains the temporal aspect of inferring clusters.

In assessing the similarities between stars, subjects may under-generalize, parsing the perceptual space more finely than the ground truth, or over-generalize, lumping stars that look dissimilar into the same galaxy (Figure 5.1). This segmentation behavior can be modeled with the prior μ subjects have about the size of perceptual space generated by each latent cause, which determines how perceptually homogeneous the stars generated from the same latent cause can be. The parameter μ gives a systematic measurement of spatial clustering.

I plan to map these three parameters onto behavioral characteristics, especially in a clinical population whose symptoms show compromised generalization. I also plan to conduct an functional magnetic resonance imaging (fMRI) study to investigate the neural mechanisms of latent cause inference.

5.3 Future Directions: scaling up to social inference

Arguably the most complex part of our environment consists of other people, whose key features such as current beliefs (Baker et al., 2017), mental states, and traits (Tamir and Thornton, 2018) are rarely observable. We often infer these unobservable states from observable actions, physical characteristics of the target person, and situational factors to make accurate predictions about behaviors of the person. How do we fathom the unfathomable mind?

The Bayesian Theory of Mind (BToM) model (Baker et al., 2017) suggests that an observer infers the inherently hidden mind (e.g., beliefs, desires, etc.) from the observable variables (e.g., actions, environments), which then guides better predictions about the target person's behaviors.

Beyond predicting other people's actions, inferring other people's hidden cognitive states (e.g., intention, knowledge) is useful when we need to learn from feedback given by other people. If we believe that someone is knowledgeable, we can make inferences about the person's knowledge from observed actions (Shafto et al., 2012). If we get feedback from someone whose intention is to teach us, we can use their feedback to infer the teacher's policy to achieve their goal to transfer as much knowledge to us, and in turn, to improve our policy (Ho et al., 2019). On top of knowledgeability and intention to teach, people consider whether the teacher is trustworthy or has bad intentions (Landrum et al., 2015). Children and adults take into account the "niceness" and "honesty" as much as the "smartness" of a teacher when deciding whom to ask for information, suggesting that we draw inferences upon traits, a deeper layer of hidden states.

Trait inference is beneficial in that it provides a stable basis for generalizing behaviors across different situations. People spontaneously form mental representations of other people's traits, even when the immediate task does not require trait inference and, in fact, could be performed better when ignoring such information. For instance, Hackel et al. (2015) had participants make a series of decisions about a partner in a game where the other player divided a pool of points between the two players. Importantly, immediate rewards (i.e., how many points the target shared with the participant on a trial) and generosity (i.e., the proportion shared from the pool) were orthogonalized, and participants could maximize their share by just focusing on the immediate rewards, ignoring the trait generosity. Supporting the idea that spontaneous trait inference guides decisions, participants used both the reward magnitude and generosity information in making their decisions, with ventral striatum responding to both reward and trait prediction errors. These results suggest the importance of having a mental model about other people and thereby building a basis that can later guide our decisions. In addition to traits like generosity, people represent how powerful people around them are relative to themselves (Kumaran et al., 2016), and the hippocampus represents information regarding other people's power and affiliation, drawing a social map that can help humans navigate the social world (Tavares et al., 2015).

We also often build a model about social groups that encompass multiple people – namely, stereotypes (Hilton and von Hippel, 1996). For example, people associate gender with different professions, and even after learning counter-stereotypic information (e.g., "Elizabeth is a doctor, and Jonathan is a nurse"), the stereotypic association lingers and slows down responses that are counter-stereotypic (Cao and Banaji, 2016). Such associations emerge early in development, biasing inferences about social groups such as racial group membership when given characteristics such as wealth (Olson et al., 2012). It is worth noting that, as in non-social decision-making, what is relevant in social decision-making can vary depending on the problem at hand. It remains an open question how inferred stable characteristics of social groups can inform state inference for stable traits and, in turn, the transient mental state of an individual.

Studying how we build models of the highly rich and "partially observable" social world is valuable not only because it would help understand social decision-making but also because it would provide deeper insights into how we represent the world as a whole. The field of social psychology has generated a wealth of research on topics that parallel questions in latent cause inference, for instance, how memory about a person and group-level representations interact with one another to guide predictions about individual behavior (Brewer et al., 1995) and how attention is allocated to information that fits our stereotypes (Plaks et al., 2001). It would be useful to draw upon this existing body of research to further our understanding of inference processes in a partially observable world.

Bibliography

Alba, J. W. and Hasher, L. (1983) Is memory schematic? *Psychological Bulletin*, **93**, 203–231.

Aldous, D. J. (1985) Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII — 1983*, vol. 1117, 1–198. Berlin, Heidelberg: Springer, Berlin, Heidelberg.

Alves, H., Koch, A. S. and Unkelbach, C. (2017a) The ”common good” phenomenon: why similarities are positive and differences are negative. *Journal of Experimental Psychology: General*.

— (2017b) Why good is more alike than bad: processing implications. *Trends in Cognitive Sciences*, **21**, 69–79.

— (2018) A cognitive-ecological explanation of intergroup biases. *Psychological Science*, **29**, 1126–1133.

Alves, H., Unkelbach, C., Burghardt, J., Koch, A. S., Krüger, T. and Becker, V. D. (2015) A density explanation of valence asymmetries in recognition memory. *Memory & Cognition*, **43**, 896–909.

Anderson, J. R. (1991) The adaptive nature of human categorization. *Psychological Review*, **98**, 409–429.

Asch, S. E. (1946) Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, **41**, 258–290.

Austerweil, J. L. and Griffiths, T. L. (2013) A nonparametric Bayesian framework for constructing flexible feature representations. *Psychological Review*, **120**, 817–851.

Baker, C. L., Jara-Ettinger, J., Saxe, R. and Tenenbaum, J. B. (2017) Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, **1**, 0064.

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U. and Norman, K. A. (2017) Discovering event structure in continuous narrative perception and memory. *Neuron*, **95**, 709–721.

Bastian, B. and Haslam, N. (2006) Psychological essentialism and stereotype endorsement. *Journal of Experimental Social Psychology*, **42**, 228–235.

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015) Fitting linear mixed-effects models using **lme4**. *Journal of Statistical Software*, **67**.

Behrens, T. E. J., Woolrich, M. W., Walton, M. E. and Rushworth, M. F. S. (2007) Learning the value of information in an uncertain world. *Nature Neuroscience*, **10**, 1214–1221.

Ben-Yakov, A. and Henson, R. N. A. (2018) The hippocampal film editor: sensitivity and specificity to event boundaries in continuous experience. *The Journal of Neuroscience*, **38**, 10057–10068.

Blei, D. M. and Frazier, P. I. (2011) Distance dependent Chinese restaurant processes. *Journal of Machine Learning Research*, **12**, 2383–2410.

Bornstein, A. M. and Norman, K. A. (2017) Reinstated episodic context guides sampling-based decisions for reward. *Nature Neuroscience*, **20**, 997–1003.

Bouton, M. E. and Bolles, R. C. (1979) Contextual control of the extinction of conditioned fear. *Learning and Motivation*, **10**, 445–466.

Brady, T. F., Schacter, D. L. and Alvarez, G. A. (2018) The adaptive nature of false memories is revealed by gist-based distortion of true memories. *PsyArXiv*.

Bransford, J. D. and Johnson, M. K. (1972) Contextual prerequisites for understanding: some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, **11**, 717–726.

Brewer, M. B., Weber, J. G. and Carini, B. (1995) Person memory in intergroup contexts: categorization versus individuation. *Journal of Personality and Social Psychology*, **69**, 29–40.

Brunec, I. K., Moscovitch, M. and Barense, M. D. (2018) Boundaries shape cognitive representations of spaces and events. *Trends in Cognitive Sciences*, **22**, 637–650.

Butz, M. V., Bilkey, D., Humaidan, D., Knott, A. and Otte, S. (2019) Learning, planning, and control in a monolithic neural event inference architecture. *Neural Networks*, **117**, 135–144.

Campbell, D. T. (1958) Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Behavioral Science*, **3**, 14–25.

Cao, J. and Banaji, M. R. (2016) The base rate principle and the fairness principle in social judgment. *Proceedings of the National Academy of Sciences*, **113**, 7475–7480.

Chen, J., Honey, C. J., Simony, E., Arcaro, M. J., Norman, K. A. and Hasson, U. (2016) Accessing real-life episodic information from minutes versus hours earlier modulates hippocampal and high-order cortical dynamics. *Cerebral Cortex*, **26**, 3428–3441.

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A. and Hasson, U. (2017) Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*, **20**, 115–125.

Collins, A. G. E. and Frank, M. J. (2013) Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychological Review*, **120**, 190–229.

Courville, A. C., Daw, N. D. and Touretzky, D. S. (2005) Similarity and discrimination in classical conditioning: a latent variable account. In *Advances in Neural Information Processing Systems 17*, 313–320.

Dayan, P. and Yu, A. J. (2006) Phasic norepinephrine: A neural interrupt signal for unexpected events. *Network: Computation in Neural Systems*, **17**, 335–350.

Deese, J. (1959) On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, **58**, 17–22.

Dehaene, S., Dehaene-Lambertz, G. and Cohen, L. (1998) Abstract representations of numbers in the animal and human brain. *Trends in Neurosciences*, **21**, 355–361.

Denrell, J. (2005) Why most people disapprove of me: experience sampling in impression formation. *Psychological Review*, **112**, 951–978.

Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D. and Daw, N. D. (2015) Model-based choices involve prospective neural activity. *Nature Neuroscience*, **18**, 767–772.

DuBrow, S. and Davachi, L. (2013) The influence of context boundaries on memory for the sequential order of events. *Journal of Experimental Psychology: General*, **142**, 1277–1286.

— (2014) Temporal memory is shaped by encoding stability and intervening item reactivation. *The Journal of Neuroscience*, **34**, 13998–14005.

— (2016) Temporal binding within and across events. *Neurobiology of Learning and Memory*, **134**, 107–114.

DuBrow, S., Rouhani, N., Niv, Y. and Norman, K. A. (2017) Does mental context drift or shift? *Current Opinion in Behavioral Sciences*, **17**, 141–146.

Duncan, K., Semmler, A. and Shohamy, D. (2019) Modulating the Use of Multiple Memory Systems in Value-based Decisions with Contextual Novelty. *Journal of Cognitive Neuroscience*, **31**, 1455–1467.

Dunsmoor, J. E., Ahs, F., Zielinski, D. J. and LaBar, K. S. (2014) Extinction in multiple virtual reality contexts diminishes fear reinstatement in humans. *Neurobiology of Learning and Memory*, **113**, 157–164.

Eich, E. (1985) Context, memory, and integrated item/context imagery. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **11**, 764–770.

Eich, J. E., Weingartner, H., Stillman, R. C. and Gillin, J. C. (1975) State-dependent accessibility of retrieval cues in the retention of a categorized list. *Journal of Verbal Learning and Verbal Behavior*, **14**, 408–417.

Eichenbaum, H. (2004) Hippocampus: cognitive processes and neural representations that underlie declarative memory. *Neuron*, **44**, 109–120.

Eyal, T., Hoover, G. M., Fujita, K. and Nussbaum, S. (2011) The effect of distance-dependent construals on schema-driven impression formation. *Journal of Experimental Social Psychology*, **47**, 278–281.

Ezzyat, Y. and Davachi, L. (2011) What constitutes an episode in episodic memory? *Psychological Science*, **22**, 243–252.

— (2014) Similarity breeds proximity: pattern similarity within and across contexts is related to later mnemonic judgments of temporal proximity. *Neuron*, **81**, 1179–1189.

Fearnhead, P. (2004) Particle filters for mixture models with an unknown number of components. *Statistics and Computing*, **14**, 11–21.

Fernandez, A. and Glenberg, A. M. (1985) Changing environmental context does not reliably affect memory. *Memory & Cognition*, **13**, 333–345.

Fiske, S. T. (1980) Attention and weight in person perception: the impact of negative and extreme behavior. *Journal of Personality and Social Psychology*, **38**, 889–906.

Fiske, S. T. and Neuberg, S. L. (1990) A Continuum of Impression Formation, from Category-Based to Individuating Processes: Influences of Information and Motivation on Attention and Interpretation. In *Advances in Experimental Social Psychology* (ed. M. P. Zanna), vol. 23, 1–74. Academic Press.

Foster, D. J. and Wilson, M. A. (2007) Hippocampal theta sequences. *Hippocampus*, **17**, 1093–1099.

Fox, E. B., Sudderth, E. B., Jordan, M. I. and Willsky, A. S. (2011) A sticky HDP-HMM with application to speaker diarization. *The Annals of Applied Statistics*, **5**, 1020–1056.

Franklin, N., Norman, K. A., Ranganath, C., Zacks, J. M. and Gershman, S. J. (2019) Structured event memory: a neuro-symbolic model of event cognition. *bioRxiv*, 541607.

Gershman, S. J., Blei, D. M. and Niv, Y. (2010) Context, learning, and extinction. *Psychological Review*, **117**, 197–209.

Gershman, S. J. and Niv, Y. (2010) Learning latent structure: carving nature at its joints. *Current Opinion in Neurobiology*, **20**, 251–256.

Gershman, S. J., Radulescu, A., Norman, K. A. and Niv, Y. (2014) Statistical computations underlying the dynamics of memory updating. *PLoS Computational Biology*, **10**, e1003939.

Ghosh, S., Ungureanu, A. B., Sudderth, E. B. and Blei, D. M. (2011) Spatial distance dependent Chinese restaurant processes for image segmentation. *Advances in Neural Information Processing Systems 24*, 1476–1484.

Gilboa, A. and Marlatte, H. (2017) Neurobiology of schemas and schema-mediated memory. *Trends in Cognitive Sciences*, **21**, 618–631.

Godden, D. R. and Baddeley, A. D. (1975) Context-dependent memory in two natural environments: on land and underwater. *British Journal of Psychology*, **66**, 325–331.

Graves, A., Fernández, S., Gomez, F. and Schmidhuber, J. (2006) Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning - ICML '06*, 369–376. Pittsburgh, PA.

Gumbsch, C., Butz, M. V. and Martius, G. (2019) Autonomous Identification and Goal-Directed Invocation of Event-Predictive Behavioral Primitives. *IEEE Transactions on Cognitive and Developmental Systems*, 1–1.

Gupta, A. S., Van Der Meer, M. A. A., Touretzky, D. S. and Redish, A. D. (2012) Segmentation of spatial experience by hippocampal theta sequences. *Nature Neuroscience*, **15**, 1032–1039.

Hackel, L. M., Doll, B. B. and Amodio, D. M. (2015) Instrumental learning of traits versus rewards: Dissociable neural correlates and effects on choice. *Nature Neuroscience*, **18**, 1233–1235.

Hamilton, D. L. and Sherman, S. J. (1996) Perceiving persons and groups. *Psychological Review*, **103**, 336–355.

Hard, B. M., Meyer, M. and Baldwin, D. (2019) Attention reorganizes as structure is detected in dynamic action. *Memory & Cognition*, **47**, 17–32.

Hard, B. M., Recchia, G. and Tversky, B. (2011) The shape of action. *Journal of Experimental Psychology: General*, **140**, 586–604. Place: US Publisher: American Psychological Association.

Hasson, U., Yang, E., Vallines, I., Heeger, D. J. and Rubin, N. (2008) A hierarchy of temporal receptive windows in human cortex. *The Journal of Neuroscience*, **28**, 2539–2550.

Hertwig, R., Barron, G., Weber, E. U. and Erev, I. (2004) Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, **15**, 534–539.

Heusser, A. C., Ezzyat, Y., Shiff, I. and Davachi, L. (2018) Perceptual boundaries cause mnemonic trade-offs between local boundary processing and across-trial associative binding. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **44**, 1075–1090.

Hilton, J. L. and von Hippel, W. (1996) Stereotypes. *Annual Review of Psychology*, **47**, 237–271.

Ho, M. K., Cushman, F., Littman, M. L. and Austerweil, J. L. (2019) People teach with rewards and punishments as communication, not reinforcements. *Journal of Experimental Psychology: General*, **148**, 520–549.

Horner, A. J., Bisby, J. A., Wang, A., Bogus, K. and Burgess, N. (2016) The role of spatial boundaries in shaping long-term event representations. *Cognition*, **154**, 151–164.

Hoskin, A. N., Bornstein, A. M., Norman, K. A. and Cohen, J. D. (2019) Refresh my memory: episodic memory reinstatements intrude on working memory maintenance. *Cognitive, Affective, & Behavioral Neuroscience*, **19**, 338–354.

Howard, M. W. and Kahana, M. J. (2002) A distributed representation of temporal context. *Journal of Mathematical Psychology*, **46**, 269–299.

Hunt, R. R. (1995) The subtlety of distinctiveness: what von Restorff really did. *Psychonomic Bulletin & Review*, **2**, 105–112.

Isarida, T. and Isarida, T. K. (2007) Environmental context effects of background color in free recall. *Memory and Cognition*, **35**, 1620–1629.

Ito, T. A. and Cacioppo, J. T. (2010) Variations on a human universal: Individual differences in positivity offset and negativity bias. *Cognition and Emotion*, **19**, 1–26.

Jeunehomme, O. and D'Argembeau, A. (2018) *Event segmentation and the temporal compression of experience in episodic memory.* Springer Berlin Heidelberg.

Johnson, A. and Redish, A. D. (2007) Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *The Journal of Neuroscience*, **27**, 12176–12189.

Jones, E. E. and Davis, K. E. (1965) From acts to dispositions: the attribution process in person perception. In *Advances in Experimental Social Psychology*, vol. 2, 219–266. Academic Press.

Kahneman, D. and Tversky, A. (1979) Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, **47**, 263–291.

Knowlton, B. J., Mangels, J. A. and Squire, L. R. (1996) A neostriatal habit learning system in humans. *Science*, **273**, 1399–402.

Kosie, J. E. and Baldwin, D. (2019) Attentional profiles linked to event segmentation are robust to missing information. *Cognitive Research: Principles and Implications*, **4**, 8.

Kumaran, D., Banino, A., Blundell, C., Hassabis, D. and Dayan, P. (2016) Compuations underlying social hierarchy learning: distinct neural mechanisms for updating and representing self-relevant information. *Neuron*, **92**, 1135–1147.

Kurby, C. A. and Zacks, J. M. (2008) Segmentation in the perception and memory of events. *Trends in Cognitive Sciences*, **12**, 72–79.

Landrum, A. R., Eaves, B. S. and Shafto, P. (2015) Learning to trust and trusting to learn: a theoretical framework. *Trends in Cognitive Sciences*, **19**, 109–111.

Lerner, Y., Honey, C. J., Silbert, L. J. and Hasson, U. (2011) Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *The Journal of Neuroscience*, **31**, 2906–2915.

Levy, S. R., Stroessner, S. J. and Dweck, C. S. (1998) Stereotype formation and endorsement: the role of implicit theories. *Journal of Personality and Social Psychology*, **74**, 1421–1436.

Lickel, B., Hamilton, D. L., Wieczorkowska, G., Lewis, A., Sherman, S. J. and Uhles, A. N. (2000) Varieties of groups and the perception of group entitativity. *Journal of Personality and Social Psychology*, **78**, 223–246.

Lieder, F., Griffiths, T. L. and Hsu, M. (2018) Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological Review*, **125**, 1–32.

Lieder, F., Hsu, M. and Griffiths, T. L. (2014) The high availability of extreme events serves resource-rational decision-making. In *Proceedings of the Annual Meeting of the Cognitive Science Society, 36*.

Lositsky, O., Chen, J., Toker, D., Honey, C. J., Shvartsman, M., Poppenk, J. L., Hasson, U. and Norman, K. A. (2016) Neural pattern change during encoding of a narrative predicts retrospective duration estimates. *eLife*, **5**, e16070.

Ludvig, E. A. and Spetch, M. L. (2011) Of black swans and tossed coins: is the description-experience gap in risky choice limited to rare events? *PLoS ONE*, **6**, e20262.

Madan, C. R., Ludvig, E. A. and Spetch, M. L. (2014) Remembering the best and worst of times: Memories for extreme outcomes bias risky decisions. *Psychonomic Bulletin and Review*.

— (2017) The role of memory in distinguishing risky decisions from experience and description. *Quarterly Journal of Experimental Psychology*, **70**, 2048–2059.

Magliano, J. P., Radvansky, G. A., Forsythe, J. C. and Copeland, D. E. (2014) Event segmentation during first-person continuous events. *Journal of Cognitive Psychology*, **26**, 649–661.

McNerney, M. W., Goodwin, K. A. and Radvansky, G. A. (2011) A novel study: a situation model analysis of reading times. *Discourse Processes*, **48**, 453–474.

Medin, D. L. and Schaffer, M. M. (1978) Context theory of classification learning. *Psychological Review*, **85**, 207–238.

Mende-Siedlecki, P., Cai, Y. and Todorov, A. (2013) The neural dynamics of updating person impressions. *Social Cognitive and Affective Neuroscience*, **8**, 623–631.

Merten, K. and Nieder, A. (2008) Compressed Scaling of Abstract Numerosity Representations in Adult Humans and Monkeys. *Journal of Cognitive Neuroscience*, **21**, 333–346.

Michelmann, S., Staresina, B. P., Bowman, H. and Hanslmayr, S. (2019) Speed of time-compressed forward replay flexibly changes in human episodic memory. *Nature Human Behaviour*, **3**, 143–154.

Momennejad, I., Otto, A. R., Daw, N. D. and Norman, K. A. (2018) Offline replay supports planning in human reinforcement learning. *eLife*, **7**, e32548.

Murnane, K., Phelps, M. P. and Malmberg, K. (1999) Context-dependent recognition memory: The ICE theory. *Journal of Experimental Psychology: General*, **128**, 403–415.

Murty, V. P., FeldmanHall, O., Hunter, L. E., Phelps, E. A. and Davachi, L. (2016) Episodic memories predict adaptive value-based decision-making. *Journal of Experimental Psychology: General*, **145**, 548–558.

Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasly, B. and Gold, J. I. (2012) Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature Neuroscience*, **15**, 1040–1046.

Nassar, M. R., Wilson, R. C., Heasly, B. and Gold, J. I. (2010) An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *The Journal of Neuroscience*, **30**, 12366–12378.

Navajas, J., Hindocha, C., Foda, H., Keramati, M., Latham, P. E. and Bahrami, B. (2017) The idiosyncratic nature of confidence. *Nature Human Behaviour*, **1**, 810–818.

Newtson, D. (1973) Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology*, **28**, 28–38.

Nosofsky, R. M. (1986) Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, **115**, 39–57.

Olson, K. R., Shutts, K., Kinzler, K. D. and Weisman, K. G. (2012) Children associate racial groups with wealth: evidence from South Africa. *Child Development*, **83**, 1884–1899.

O'Reilly, J. X. (2013) Making predictions in a changing world-inference, uncertainty, and learning. *Frontiers in Neuroscience*, **7**, 105.

Pearce, J. M. and Hall, G. (1980) A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, **87**, 532–552.

Pettijohn, K. A. and Radvansky, G. A. (2016) Narrative event boundaries, reading times, and expectation. *Memory & Cognition*, **44**, 1064–1075.

Pezzulo, G., van der Meer, M. A. A., Lansink, C. S. and Pennartz, C. M. A. (2014) Internally generated sequences in learning and executing goal-directed behavior. *Trends in Cognitive Sciences*, **18**, 647–657.

Pfeiffer, B. E. and Foster, D. J. (2013) Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*, **497**, 74–79.

Plaks, J. E., Stroessner, S. J., Dweck, C. S. and Sherman, J. W. (2001) Person theories and attention allocation: preferences for stereotypic versus counterstereotypic information. *Journal of Personality and Social Psychology*, **80**, 876–893.

Poldrack, R. A., Clark, J. J., Pare-Blagoev, E. J., Shohamy, D., Moyano, J. C., Myers,

C. E. and Gluck, M. A. (2001) Interactive memory systems in the human brain. *Nature*, **414**, 546–550.

Polyn, S. M., Norman, K. A. and Kahana, M. J. (2009) Task context and organization in free recall. *Neuropsychologia*, **47**, 2158–2163.

Poppenk, J. L. and Norman, K. A. (2012) Mechanisms supporting superior source memory for familiar items: a multi-voxel pattern analysis study. *Neuropsychologia*, **50**, 3015–3026.

Radvansky, G. A. (2012) Across the event horizon. *Current Directions in Psychological Science*, **21**, 269–272.

Radvansky, G. A. and Copeland, D. E. (2006) Walking through doorways causes forgetting: situation models and experienced space. *Memory & Cognition*, **34**, 1150–1156.

— (2010) Reading times and the detection of event shift processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **36**, 210–216.

Radvansky, G. A., Krawietz, S. A. and Tamplin, A. K. (2011) Walking through doorways causes forgetting: further explorations. *Quarterly Journal of Experimental Psychology*, **64**, 1632–1645.

Radvansky, G. A. and Zacks, J. M. (2011) Event perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, **2**, 608–620.

— (2017) Event boundaries in memory and cognition. *Current Opinion in Behavioral Sciences*, **17**, 133–140.

Radvansky, G. A. and Zacks, R. T. (1991) Mental models and the fan effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **17**, 940–953.

Radvansky, G. A., Zwaan, R. A., Federico, T. and Franklin, N. (1998) Retrieval from temporally organized situation models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **24**, 1224–1237.

Reed, S. K. (1972) Pattern recognition and categorization. *Cognitive Psychology*, **3**, 382–407.

Reggente, N., Essoe, J. K.-Y., Aghajan, Z. M., Tavakoli, A. V., McGuire, J. F., Suthana, N. A. and Rissman, J. (2018) Enhancing the ecological validity of fMRI memory research using virtual reality. *Frontiers in Neuroscience*, **12**, 408.

von Restorff, H. (1933) Über die Wirkung von Bereichsbildungen im Spurenfeld. *Psychologische Forschung*, **18**, 299–342.

Reynolds, J. R., Zacks, J. M. and Braver, T. S. (2007) A computational model of event segmentation from perceptual prediction. *Cognitive Science*, **31**, 613–643.

Richmond, L. L. and Zacks, J. M. (2017) Constructing experience: event models from perception to action. *Trends in Cognitive Sciences*, **21**, 962–980.

Rinck, M. and Weber, U. (2003) Who when where: an experimental test of the event-indexing model. *Memory & Cognition*, **31**, 1284–1292.

Roediger, H. L. and McDermott, K. B. (1995) Creating false memories: remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **21**, 803–814.

Rouhani, N., Norman, K. A. and Niv, Y. (2018) Dissociable effects of surprising rewards on learning and memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **44**, 1430–1443.

Rouhani, N., Norman, K. A., Niv, Y. and Bornstein, A. M. (2020) Reward prediction errors create event boundaries in memory. *Cognition*, **203**, 104269.

Rozin, P. and Royzman, E. B. (2001) Negativity bias, negativity dominance, and dontagion. *Personality and Social Psychology Review*, **5**, 296–320.

Saffran, J. R., Aslin, R. N. and Newport, E. L. (1996) Statistical Learning by 8-Month-Old Infants. *Science*, **274**, 1926–1928.

Sanborn, A. N., Griffiths, T. L. and Navarro, D. J. (2010) Rational approximations to rational models: alternative algorithms for category learning. *Psychological Review*, **117**, 1144–1167.

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B. and Botvinick, M. M. (2013) Neural representations of events arise from temporal community structure. *Nature Neuroscience*, **16**, 486–492.

Schuck, N. W. and Niv, Y. (2019) Sequential replay of nonspatial task states in the human hippocampus. *Science*, **364**, 1254.

Shadlen, M. N. and Shohamy, D. (2016) Decision making and sequential sampling from memory. *Neuron*, **90**, 927–939.

Shafto, P., Goodman, N. D. and Frank, M. C. (2012) Learning from others: the consequences of psychological reasoning for human learning. *Perspectives on Psychological Science*, **7**, 341–351.

Shepard, R. N. (1987) Toward a universal law of generalization for psychological science. *Science*, **237**, 1317–1323.

Shohamy, D. and Daw, N. D. (2015) Integrating memories to guide decisions. *Current Opinion in Behavioral Sciences*, **5**, 85–90.

Siefke, B. M., Smith, T. A. and Sederberg, P. B. (2019) A context-change account of temporal distinctiveness. *Memory & Cognition*, **79**, 1–15.

Siegel, J. Z., Mathys, C., Rutledge, R. B. and Crockett, M. J. (2018) Beliefs about bad people are volatile. *Nature Human Behaviour*, **2**, 750–756.

Smith, E. R. and Zárate, M. A. (1992) Exemplar-based model of social judgment. *Psychological Review*, **99**, 3–21.

Smith, S. M. (1979) Remembering in and out of context. *Journal of Experimental Psychology: Human Learning and Memory*, **5**, 460–471.

Smith, S. M. and Manzano, I. (2010) Video context-dependent recall. *Behavior Research Methods*, **42**, 292–301.

Smith, S. M. and Vela, E. (2001) Environmental context-dependent memory: a review and meta-analysis. *Psychonomic Bulletin & Review*, **8**, 203–220.

Socher, R., Gershman, S. J., Perotte, A. J., Sederberg, P. B., Blei, D. M. and Norman, K. A. (2009) A bayesian analysis of dynamics in free recall. In *Advances in Neural Information Processing Systems 22*, 1714–1722.

Soderstrom, N. C. and McCabe, D. P. (2011) Are survival processing memory advantages based on ancestral priorities? *Psychonomic Bulletin & Review*, **18**, 564–569.

Soto, F. A., Gershman, S. J. and Niv, Y. (2014) Explaining compound generalization in associative and causal learning through rational principles of dimensional generalization. *Psychological Review*, **121**, 526–558.

Speer, N. K., Swallow, K. M. and Zacks, J. M. (2003) Activation of human motion processing areas during event perception. *Cognitive, Affective and Behavioral Neuroscience*, **3**, 335–345.

Speer, N. K. and Zacks, J. M. (2005) Temporal changes as event boundaries: processing and memory consequences of narrative time shifts. *Journal of Memory and Language*, **53**, 125–140.

Speer, N. K., Zacks, J. M. and Reynolds, J. R. (2007) Human brain activity time-locked to narrative event boundaries: Research article. *Psychological Science*, **18**, 449–455.

Swallow, K. M., Barch, D. M., Head, D., Maley, C. J., Holder, D. and Zacks, J. M. (2011) Changes in events alter how people remember recent information. *Journal of Cognitive Neuroscience*, **23**, 1052–1064.

Swallow, K. M., Zacks, J. M. and Abrams, R. A. (2009) Event boundaries in perception affect memory encoding and updating. *Journal of Experimental Psychology: General*, **138**, 236–257.

Tamir, D. I. and Thornton, M. A. (2018) Modeling the predictive social mind. *Trends in Cognitive Sciences*, **22**, 201–212.

Tavares, R. M., Mendelsohn, A., Grossman, Y., Williams, C. H., Shapiro, M., Trope, Y. and Schiller, D. (2015) A map for social navigation in the human brain. *Neuron*, **87**, 231–43.

Tenenbaum, J. B. and Griffiths, T. L. (2001) Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, **24**, 629–640.

Tse, D., Langston, R. F., Kakeyama, M., Bethus, I., Spooner, P. A., Wood, E. R., Witter, M. P. and Morris, R. G. M. (2007) Schemas and memory consolidation. *Science*, **316**, 76–82.

Tulving, E. and Thomson, D. M. (1973) Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, **80**, 352–373.

Underwood, G. (1977) Contextual facilitation from attended and unattended messages. *Journal of Verbal Learning and Verbal Behavior*, **16**, 99–106.

Unkelbach, C., Fiedler, K., Bayer, M., Stegmüller, M. and Danner, D. (2008) Why positive information is processed faster: the density hypothesis. *Journal of Personality and Social Psychology*, **95**, 36–49.

Watkins, M. J. and Watkins, O. C. (1976) Cue-overload theory and the method of interpolated attributes. *Bulletin of the Psychonomic Society*, **7**, 289–291.

Weiss, W. and Margolius, G. (1954) The effect of context stimuli on learning and retention. *Journal of Experimental Psychology*, **48**, 318–322.

Whittington, J. C., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N. and Behrens, T. E. (2019) The Tolman-Eichenbaum machine: unifying space and relational memory through generalisation in the hippocampal formation. *bioRxiv*, 770495.

Wälti, M. J., Woolley, D. G. and Wenderoth, N. (2019) Reinstating verbal memories with virtual contexts: Myth or reality? *PLOS ONE*, **14**, e0214540.

Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., Buckner, R. L. and Raichle, M. E. (2001a) Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience*, **4**, 651–655.

Zacks, J. M., Kurby, C. A., Eisenberg, M. L. and Haroutunian, N. (2011) Prediction error associated with the perceptual segmentation of naturalistic events. *Journal of Cognitive Neuroscience*, **23**, 4057–4066.

Zacks, J. M., Speer, N. K. and Reynolds, J. R. (2009) Segmentation in reading and film comprehension. *Journal of Experimental Psychology: General*, **138**, 307–327.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S. and Reynolds, J. R. (2007) Event perception: a mind-brain perspective. *Psychological Bulletin*, **133**, 273–293.

Zacks, J. M., Tversky, B. and Iyer, G. (2001b) Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General*, **130**, 29–58.

Zhang, H. and Maloney, L. T. (2012) Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience*, **6**, 1.

Zwaan, R. A. (1996) Processing narrative time shifts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **22**, 1196–1207.

Zwaan, R. A., Langston, M. C. and Graesser, A. C. (1995a) The construction of situation models in narrative comprehension: an event-indexing model. *Psychological Science*, **6**, 292–297.

Zwaan, R. A., Magliano, J. P. and Graesser, A. C. (1995b) Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **21**, 386–397.

Zwaan, R. A. and Radvansky, G. A. (1998) Situation models in language comprehension and memory. *Psychological Bulletin*, **123**, 162–185.

Zwaan, R. A., Radvansky, G. A., Hilliard, A. E. and Curiel, J. M. (1998) Constructing multidimensional situation models during reading. *Scientific Studies of Reading*, **2**, 199–220.

Ólafsdóttir, H. F., Bush, D. and Barry, C. (2018) The role of hippocampal replay in memory and planning. *Current Biology*, **28**, R37–R50.